Studies in Educational Evaluation 54 (2017) 22–42

Contents lists available at ScienceDirect

Studies in Educational Evaluation

journal homepage: www.elsevier.com/stueduc

Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related

Bob Uttl*, Carmela A. White[1], Daniela Wong Gonzalez[2]

Department of Psychology, Mount Royal University, Canada

ARTICLE INFO

Article history:
Received 23 February 2016
Received in revised form 10 August 2016
Accepted 17 August 2016
Available online 19 September 2016

Keywords:
Meta-analysis of student evaluation of teaching
Multisection studies
Validity
Teaching effectiveness
Evaluation of faculty
SET and learning correlations

ABSTRACT

Student evaluation of teaching (SET) ratings are used to evaluate faculty's teaching effectiveness based on a widespread belief that students learn more from highly rated professors. The key evidence cited in support of this belief are meta-analyses of multisection studies showing small-to-moderate correlations between SET ratings and student achievement (e.g., Cohen, 1980, 1981; Feldman, 1989). We re-analyzed previously published meta-analyses of the multisection studies and found that their findings were an artifact of small sample sized studies and publication bias. Whereas the small sample sized studies showed large and moderate correlation, the large sample sized studies showed no or only minimal correlation between SET ratings and learning. Our up-to-date meta-analysis of all multisection studies revealed no significant correlations between the SET ratings and learning. These findings suggest that institutions focused on student learning and career success may want to abandon SET ratings as a measure of faculty's teaching effectiveness.

"*For every complex problem there is an answer that is clear, simple, and wrong.*" H. L. Mencken

Student Evaluation of Teaching (SET) ratings are used to evaluate faculty's teaching effectiveness based on an assumption that students learn more from highly rated professors. Although SET were used as early as 1920's, their use expanded across the USA in the late 1960's and early 1970's (Murray, 2005; Wachtel, 1998). Today, nearly all colleges and universities in Norh America use SET to evaluate their faculty's teaching effectiveness (Murray, 2005; Wachtel, 1998). Typically, SET are conducted within the last few weeks of courses, before the final grades are assigned. Students are presented with rating forms that ask them to rate their perceptions of instructors and courses, often on a 5-point Likert scale ranging from Strongly Disagree to Strongly Agree. The rating forms may ask students to provide overall ratings of instructor and/or course and they may also ask students to rate numerous specific characteristics of teachers (e.g., knowledge, clarity of explanation, organization, enthusiasm, friendliness, fairness, availability, approachability, use of humor, contribution to students' learning) and courses (e.g., organization, difficulty) (Feldman, 1989; Spooren, Brockx, & Mortelmans, 2013). The ratings for each course/class are summarized, typically by calculating mean ratings across all responding students for each rated item and across all rated items, and these mean class SET ratings are then used to evaluate professors' teaching effectiveness by comparing them, for example, to department or university average ratings. Although use of SET as a feedback for professors' own use is not controversial, the use of SET as a measure of professors' teaching effectiveness for making high stakes administrative decisions about instructors' hiring, firing, merit pay, and promotions is highly controversial (e.g., Emery, Kramer, & Tian, 2003; Spooren, Brockx, & Mortelmans, 2013; Stark & Freishtat, 2014; Wachtel, 1998).

Proponents of SET as a measure of instructor teaching effectiveness have put forward a number of reasons for their use: (1) SET are cheap and convenient means to evaluate faculty's teaching, (2) SET are very useful to demonstrate administrators' concerns with public accountability and public relations, (3) SET allow students to have say in evaluation of faculty's teaching, and (4) students are uniquely positioned to evaluate their experiences and perceptions of instructors as they are teaching classes (Murray, 2005; Wachtel, 1998). The last reason on this list is the SET proponents' main rationale for why SET ought to measure instructor's teaching effectiveness. The SET proponents assume that students observe instructors' behavior, assess how much they learned from the instructor, rate the instructor according to how

* Corresponding author at: Department of Psychology, Mount Royal University, 4825 Mount Royal University Gate, Calgary, AB, T3E 6K6, Canada.
E-mail address: buttl@mtroyal.ca (B. Uttl).
[1] Department of Psychology, University of British Columbia, Okanagan, Canada.
[2] Department of Psychology, University of Windsor, Canada.

http://dx.doi.org/10.1016/j.stueduc.2016.08.007
0191-491X/

much the instructor's contributed to their learning, and thus, high correlation between SET and measures of learning should follow. In contrast, the opponents of SET as measure of teaching effectiveness argue that SET are primarily measures of student satisfaction, that is, "a happy or pleased feeling because of something that you did or something that happened to you" (www.merriam-webster.com). Clearly, whether a student is overwhelmed by "happy or pleased feeling" at the end of the course is likely to depend on many factors that have nothing to do with instructor's teaching effectiveness, for example, whether or not a student was getting grades that he or she thought she deserved to be getting throughout a course, whether or not a course was forced on a student by being required, whether or not a student was reported by an instructor for cheating or plagiarism, whether or not a student found instructor's accent or looks pleasant, etc. The opponents of SET as measures of teaching effectiveness argue that SET have no or only limited validity as a measure of instructor teaching effectiveness because both SET and measures of learning are influenced by teaching effectiveness irrelevant factors (TEIFs) such as academic discipline/field of study, student interest, student motivation, instructor sex, instructor accent, class level, class size, class meeting time, etc. (Franklin & Theall, 1995; Hoyt & Lee, 2003; Spooren, Brockx, & Mortelmans, 2013; Uttl et al., 2012; Wachtel, 1998). Although thousands of studies have examined validity of SET, including influence of various TEIFs on SETs, the gulf between the proponents and opponents of SET is as wide as ever.

However, the well established findings in cognitive psychology and intelligence literature suggest that any substantive correlations between SET and learning are likely to be a fluke or an artifact rather than due to students' ability to accurately assess instructor' teaching effectiveness. First, how well students do on measures of learning is dependent to large degree on students' intelligence or ability to learn, prior relevant knowledge, and motivation to learn. Second, students' ability to judge how much they learned is also dependent on their intelligence or ability. One of the well-established findings in cognitive psychology is so called Dunning-Kruger effect (Kruger & Dunning, 1999) showing that unskilled persons asses their ability to be much higher than it really is and that highly skilled persons underestimate their ability and assume that tasks they found easy were also easy for others. In one set of studies, Dunning and Kruger examined Cornell University students' self-assessment of logical reasoning skills, grammatical skills, and humor. When the students were showed their scores and asked to estimate their own rank in the class, the competent students estimated their rank accurately whereas the incompetent students overestimated their rank to such a degree that they believed their work deserved B or better grade even though their work received Ds or Fs.

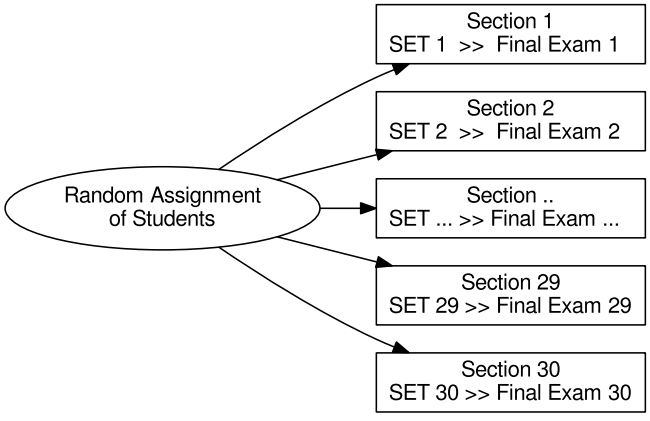

Fig. 1. Multisection SET validity study design.

The key evidence cited in support of the belief that SET measure instructor's teaching effectiveness are multisection studies showing correlations between SETs and student achievements, the correlations that have been acknowledged and accepted as true by both proponents and opponents of SETs. Fig. 1 describes the logic of multisection studies. An ideal multisection study design includes the following features: a course has many equivalent sections following the same outline and having the same assessments, students are randomly assigned to sections, each section is taught by a different instructor, all instructors are evaluated using SETs at the same time and before a final exam, and student learning is assessed using the same final exam. If students learn more from more highly rated professors, sections' average SET ratings and sections' average final exam scores should be positively correlated. However, random assignment of students to sections is rarely possible. Accordingly, some multisection studies control for prior learning/ability by, for example, regressing individual students' achievement scores on measures of students' prior learning/ability and using residual gains in achievement, averaged across all students within sections, as measures of achievement/learning. In general, researchers have agreed that multisection study designs are the best for determining the relationship between SET ratings and student learning facilitated by professors. For example, Abrami, d'Appolonia, and Cohen (1990) summarize this view as follows: "The multisection validation design is the strongest design for assessing the degree to which student ratings predict teacher-produced student learning" (p. 230).

More than three decades ago, Cohen (1981) conducted the first meta-analysis of multisection studies and reported that SET ratings correlate with student learning with $r = .43$, a small-to-moderate correlation. Cohen wrote: "The results of the meta-analysis provide strong support for the validity of student ratings as a measure of teaching effectiveness" (p. 281) and " . . . we can safely say that student ratings of instruction are a valid index of instructional effectiveness. Students do a pretty good job of distinguishing among teachers on the basis of how much they have learned" (p. 305). Since that time, Cohen's meta-analysis has been frequently cited hundreds of times in support of using SETs to evaluate faculty's teaching effectiveness (see Table 1) and the view that multisection studies have demonstrated validity of SET ratings as a measure of teaching effectiveness – that students learn more from highly rated professors – has been accepted as the established fact in various research summaries and widely disseminated to faculty members, administrators, and general public. Even self-help books designed to improve teaching of beginning faculty members inform them that the research has established that SET ratings measure teaching effectiveness. A few quotes from these reviews and materials will suffice:

> "We (d'Apollonia & Abrami, 1977) reviewed the research evidence from multisection validity studies and found that ratings explain instructor impacts on student learning to a moderate extent (corrected $r = .47$)" (Abrami & d'Apollonia, 1999, p. 519)

> "Overall, multisection validity studies have shown substantial correlation with student achievement as measured by examination performance. (Abrami, d'Apollonia, and Cohen, 1990; Abrami and d'Apollonia, 1997)" (Ory & Ryan, 2001, p. 43)

> " . . . meta-analyses of multisection validity research have supported the validity of SETs by demonstrating that the sections that evaluate the teaching as most effective are also the sections that perform best on standardized final examinations (Cohen, 1981, 1987; Feldman, 1989)· · ·This research demonstrates that teachers who receive better SETs are also the

Table 1
Key features and findings of previous meta-analyses.

	Co81	Co82	Do82	Co83	Mc84	Fe89	Cl09
Search replicable	No	No	No	No	No	No	No
Individual studies identified	No	No	No	No	No	No	No
Characteristics of individual studies described	No	No	Yes	Yes	No	Yes	Yes
Number of articles	41	16	5	22	14	32	17
Number of studies	68	21	5	40	42*	48	42
Size of each study (#sections) reported	No	No	Yes	Yes	Yes	Yes	Yes
Effect size for each study reported	No	No	Yes	Yes	Yes	Yes	Yes
Weighted estimates of *r*s used	No	?	Yes	Yes	Yes	No	Yes
Fisher's Z transformed averages for *r*s reported	Yes	Yes	Yes	Yes	Yes	Yes	No
Simple averages for *r*s reported	No	No	No	No	No	Yes	Yes
Scatterplot of effect size by study size reported	No	No	No	No	No	No	No
Small *n* bias considered	No	No	No	No	No	No	No
Outliers and their influence considered	No	No	No	No	No	No	No
Overall instructor SET/learning *r*	.43	.44	–	.38	.32	–	.13
Overall course SET/learning *r*	.47	.48	–	–	.26	–	–
Overall SET/learning *r* (across all SET items)	–	–	.26	–	–	–	–
Individual SET dimensions *r* minimum	−.02	-.05	–	–	–	.07	–
Individual SET dimensions *r* maximum	.50	.50	–	–	–	.57	–
Number of dimensions examined	8	6	–	–	–	31	–
Number of citations in Web of Science	253	16	10	–	23	106	–
Number of citations in Google Scholar	949	48	30	62	67	427	168

Note. Co81 = Cohen (1981), Co82 = Cohen (1982), Do82 = Dowell and Neal (1982), Co83 = Cohen (1983), Mc84 = McCallum (1984), Fe89 = Feldman (1989), Cl09 = Clayson (2009).

> teachers from whom students learn the most. Perhaps more than any other area of SET research, results based on the multisection validity paradigm support the validity of SETs." (Marsh, 2007, p. 339)
>
> " . . . the multisection studies show that classes in which the students gave the instructor higher ratings tended to be the ones where the students learned more (i.e., scored higher on the external exam)." (Benton & Cashin, 2012, p. 4)
>
> "Ratings of overall teaching effectiveness are moderately correlated with independent measures of student learning and achievement. Students of highly rated teachers achieve higher final exam scores, can better apply course material, and are more inclined to pursue the subject subsequently." (Davis, 2009, p. 534)

However, the most recent meta-analysis of the multisection studies by Clayson (2009) concluded that SET ratings are not related to student learning. Specifically, Clayson reported that the correlation between SET and learning was only .33 when correlations reported in the primary studies were averaged regardless of the sample size and only .13 when they were weighted by the sample size. What is the reason for this discrepancy in findings? In an attempt to explain the discrepancy, Clayson (2009) argued that the validity of SET may depend on "faculty, class topic matter, and academic disciplines," and dismissed possible methodological flaws with Cohen's (1981) research as an explanation. However, a quick examination of Clayson's meta-analysis data (see Clayson's Table 1) shows impossibly high (e.g., .91, .89, .81) correlations between SETs and learning reported in a number of primary studies. Moreover, Clayson's data reveal several striking observations: (1) the SET/learning correlation shrunk from .33 to .13 when he weighted the correlations by the sample size; (2) studies with a few sections reported the highest correlations whereas studies with many sections reported smaller correlations; and (3) scatterplot of the correlations against study sizes reported in Clayson's Table 1 reveals an asymmetric funnel plot, indicating small study size bias. Thus, Cohen's highly cited findings may be an artifact of his failure to notice impossibly high SET/learning correlations and to adequately consider the negative correlation between the SET/learning correlations and sample size.

A detailed review of all previously published meta-analyses of the SET/learning correlations (Clayson, 2009; Cohen, 1981, 1982, 1983; Dowell & Neal, 1982; Feldman, 1989; McCallum, 1984) reveal that none of them adequately considered the possibility that small-to-moderate SET/learning correlations may be an artifact of small sample sizes of most of the primary studies and small sample bias. The review also reveals that the previous meta-analyses suffer from multiple critical methodological flaws that render their conclusions unwarranted. The necessary, but not sufficient, first step for conducting a valid and informative meta-analysis is to gather all relevant studies and to accurately extract and report relevant information from these primary studies. A meta-analysis must, at minimum, describe the search strategies for primary studies, provide the basic descriptive information including effect size and sample size for all primary studies, and ensure that the extracted primary study level data are accurate (e.g., by assessing reliability of coding). Table 1 lists the seven previously published meta-analyses of the SET-learning correlations and highlights that all of them failed this minimum standard. Only Feldman (1989) described the repeatable search strategy: "The present analysis uses the same set of data as that used by Peter Cohen (1981, 1980, 1987)." However, Cohen's search strategy itself is not repeatable (Cohen, 1981). Not surprisingly, there are huge discrepancies in the number of studies located and the number of multisection studies extracted from those reports by different meta-analysts (see Table 1). For example, using the same inclusion/exclusion criteria when searching for the studies that employed ability/previous knowledge controls, Dowell and Neal (1982) retrieved only five articles with five multisection studies in total whereas Cohen (1983) reported that he found 22 articles with 40 multisection studies in total.

Table 1 also shows that two out of seven meta-analyses did not actually provide basic descriptive information including effect sizes and sample sizes for all primary studies (Cohen, 1981, 1982). The critical importance of providing this information is highlighted

by McCallum's (1984) meta-analysis. McCallum listed both SET/learning correlations and sample sizes he extracted from the primary studies, and thus, allowed us to verify his work and to find out that the many of the listed sample sizes and correlations were completely incorrect.

Table 1 also highlights that none of the previous meta-analyses included a simple scatterplot of correlation size as a function of number of sections (or funnel plots) and they all failed to seriously consider that the small-to-moderate correlations between SET and learning may be an artifact of small samples of most studies and small sample bias. Small sample bias can arise due to a publication bias where studies reporting statistically significant correlations are more likely to be published than those that report non significant correlations. Two meta-analyses briefly considered the effect size/sample size relationship in multisection studies. First, Cohen (1981) examined a number of moderator variables but reported that the correlation between SET/learning correlations and study sizes was not significant and only −.14 (see Cohen's Table 5, p. 300). Accordingly, Cohen concluded that number of sections is unrelated to the reported correlations. However, a few pages later, Cohen noted that "some reviewers have been concerned that rating/achievement correlations vary according to the number of sections used in the study," dismissed their concerns, but at the same time noted that "actually, number of sections correlated significantly with the *absolute value* [emphasis added] of effect size" (p. 303) without reporting the size of this correlation. Second, Clayson (2009) summarized SET/learning correlations from primary studies using both raw averages and averages weighted by the number of sections on which each correlation was based. However, he did not realize that one of the most likely explanations for the small-to-moderate SET/learning correlations was a simple statistical fact that small samples require very large correlations for statistical significance and that journals are more likely to publish significant rather than non-significant

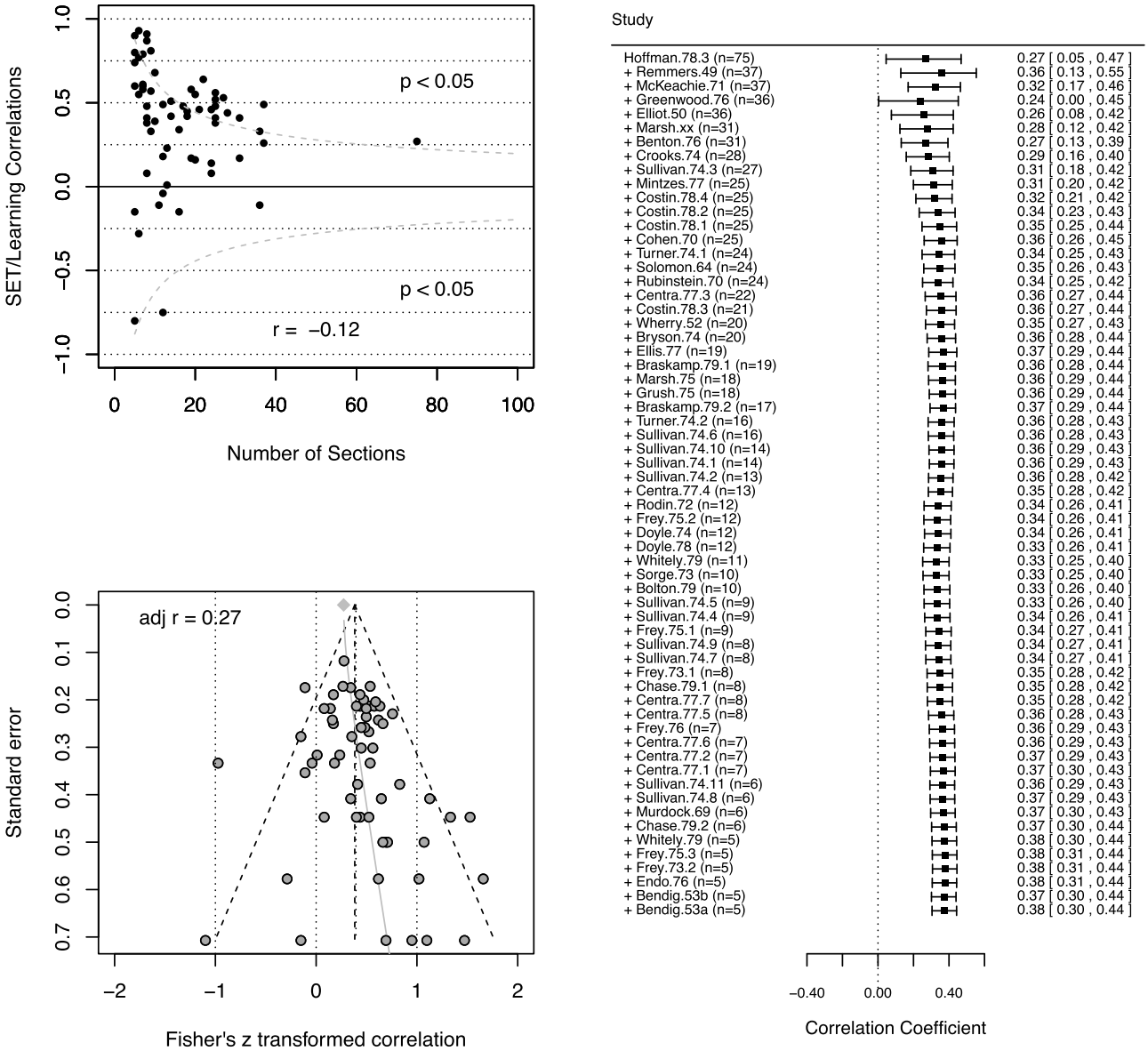

Fig. 2. Re-analysis of Cohen's (1981) multisection data set. The top left panel shows the small size study effects. The right panel shows cumulative meta-analysis showing that the meta-analysis of large sized studies reveals small SET/learning correlation and that addition of smaller size studies increases estimated SET/learning correlation. The bottom left panel shows the result of limit meta-analysis taking into account small size study effects, including adjusted $r = .27$.

Table 2
Studies included in previous and current meta-analyses including number of sections and effect sizes for each study.

Study	Course	Co81	Co82	Do82	Co83	Mc84	Fe89	Cl09	*n*	Co81 *r*	Cl09 *r*	CIS *r*	CAS *r*	Adj
Beleche et al. (2012)	Core discipline								82			.21	.16	1
Benbassat and Bachar (1981)	Clinical Medicine (intro)								15			.18	.18	1
Bendig (1953a)	Psychology (intro)	1	1			1–5		1	5	.90	.89	.89	.85	0
Bendig (1953b)	Psychology (intro)	2	2						5	−.80		−.80	−.40	0
Benton and Scott (1976)	English	3					1		31	.17		.17	.12	0
Bolton et al. (1979)	Psychology (intro)	4	3		1		2		10	.68		.68	.53	0
!!Borg and Hamilton (1956)	Military training	5				–			89	.19				
Braskamp, Caulley, and Costin (1979).01	Psychology (intro)	6	4			6	3	2	19	.17	.17	.17	.02	0
Braskamp et al. (1979).02	Psychology (intro)	7	5			7	4	3	17	.48	.48	.48	.23	0
Bryson (1974)	Algebra	8			2		5		20	.55		.55	.37	0
Capozza (1973)	Macroeconomics (i/m)								8			−.94	−.94	1
Centra (1977).01	Chemistry 1000	9			9	8	6	9	7	.60	.60	.60	.50	0
Centra (1977).02	Biology 1010	10			8	9	7	10	7	.61	.61	.61	.48	0
Centra (1977).03	Psychology 1001	11	6		7	10	8	4	22	.64	.64	.64	.34	1
Centra (1977).04	Biology 1011	12			6	11	9	5	13	.23	.23	.23	.11	1
Centra (1977).05	Math 1011	13			5	12	10	6	8	.87	.87	.87	.68	1
Centra (1977).06	Physics 1051	14			4	13	11	7	7	.58	.58	.58	.11	1
Centra (1977).07	Chemistry 1001	15			3	14	12	8	8	.41	.41	.41	.23	1
Chase et al. (1979).01	German	17			10		13		8	.93		.78	.62	1
Chase et al. (1979).02	Accounting	16			11		14		6	.38		.19	.18	1
Cohen and Berger (1970)	Natural Science	18					15		25	.48		.42	.29	0
!!Cohen (1981)	n/a							11			.41			
Costin (1978).01	Psychology (1973)	19	7			15	16	12	25	.52	.52	.52	.52	0
Costin (1978).02	Psychology (1974)	20				16	17	13	25	.56	.56	.56	.56	0
Costin (1978).03	Psychology (1975)	21				17	18	14	21	.46	.46	.46	.46	0
Costin (1978).04	Psychology (1976)	22				18	19	15	25	.41	.41	.41	.41	0
#Crooks and Smock (1974)		23			12					.44				
Doyle and Whitely (1974)	French	25		5	14		21	16	12	.49	.49	.49	.19	1
Doyle and Crichton (1978)	Communications (intro)	24			13		20		10	−.04		−.04	.02	1
Drysdale (2010).01	Algebra (intermed)								11			.09	.08	0
Drysdale (2010).02	Algebra (intermed)								10			−.02	−.11	0
Drysdale (2010).03	Algebra (intermed)								8			.64	.68	0
Drysdale (2010).04	Algebra (intermed)								11			.03	.08	0
Drysdale (2010).05	Algebra (intermed)								10			−.23	−.31	0
Drysdale (2010).06	Algebra (intermed)								12			−.10	−.08	0
Drysdale (2010).07	Algebra (intermed)								11			.19	.22	0
Drysdale (2010).08	Algebra (intermed)								16			.41	.36	0
Drysdale (2010).09	Algebra (intermed)								11			.23	.18	0
Elliott (1950)	Chemistry	27		1	15		22		36	.33		.32	.23	1
Ellis and Rickard (1977)	Psychology (intro)	26	8		16		23		19	.58		.58	.56	0
Endo and Della-Piana (1976)	Trigonometry	28					24		5	−.15		−.15	.10	0
Fenderson et al. (1997)	Pathology								29			.09	.09	0
Frey (1973).01	Calculus (intro)	29			17	19	26	17	8	.91	.91	.91	.63	1
Frey (1973).02	Multidimensional Calculus	30		3	18		25	18	5	.60	.60	.60	.60	1
Frey et al. (1975).01	Calculus (intro)	32			20	21	30	21	9	.81	.81	.81	.49	0
Frey et al. (1975).02	Educational Psychology	33	9				28	19	12	.18	.18	.18	.10	0
Frey et al. (1975).03	Calculus (intro)	34					29	20	5	.74	.74	.74	.46	0
Frey (1976)	Calculus (intro)	31			19	20	27		7	.79		.79	.41	1
Galbraith et al., (2012).01	Marketing (intro)								8			.23	.22	0
Galbraith et al., (2012).02	Macroeconomics								10			.32	.34	0
Galbraith et al., (2012).03	Finance (grad)								12			−.07	−.02	0
Galbraith et al., (2012).04	Criminal Justice (grad)								8			.31	.30	0
Galbraith et al., (2012).05	Marketing								8			−.13	−.07	0
Galbraith et al., (2012).06	Marketing (grad)								9			−.16	−.13	0
Galbraith et al., (2012).07	Statistics								13			.11	.12	0
Galbraith and Merrill (2012)	Finance (grad)								5			.29	.29	0
!!Gessner (1973).01	Basic science					22								
!!Gessner (1973).02						23								
Greenwood et al. (1976)	Analytic Geometry	35			21		31		36	−.11		−.11	−.08	1
Grush and Costin (1975)	Psychology (intro)	36	10				32		18	.45		.45	.45	0
Hoffman (1978).02	Speech	37					34		28	.27				
Hoffman (1978).03	Math (intro)	38			22		33		75	.27		.29	.25	1
!!Johnson (2003)	Various							22			−.11			
Koon and Murray (1995)	Psychology (intro)								36			.30	.30	0
Marsh et al. (1975)	Computer programming	39				24	35		18	.42		.42	.29	0
Marsh and Overall (1980)	Engineering 10	44				25	36		31	.41		.38	.36	1
McKeachie et al. (1971).01	General Psychology	40	11	2	23		37		34	.26		.06	.09	1
McKeachie et al. (1971).02	General Psychology		18						32			−.20	.06	1
McKeachie et al. (1971).03	General Psychology		19						6			.10	.01	0
McKeachie et al. (1971).04	French								16			.25	.13	0
McKeachie et al. (1971).05	Economics (intro)								18			.55	.10	0
McKeachie et al. (1978)	Psychology (intro)								6			.20	.20	0
Mintzes (1977)	Psychology (intro)	41	12				38		25	.38		.38	.30	0
Morgan and Vasché (1978)	Macroeconomics (intro)						39		5			.92	.86	1
!!Morsh et al. (1956)	Hydraulics	42			24	26	40		106	.40				

Table 2 (*Continued*)

Study	Course	Co81	Co82	Do82	Co83	Mc84	Fe89	Cl09	*n*	Co81 *r*	Cl09 *r*	CIS *r*	CAS *r*	Adj
Murdock et al. (1969)	Psychology (intro)	43	13						6	.77		.77	.77	0
#Murray (1983)							41							
Orpen (1980)	Math (intro)					27	42		10			.61	.52	0
Palmer (1978)	Microeconomics			6	25			23	14		−.16	−.17	−.17	1
Prosser and Trigwell (1991)	Nursing Communications								11			−.42	−.28	0
Rankin et al. (1965)	Developmental reading	45					43		21			−.06	.02	1
Remmers, Martin, and Elliot (1949)	Chemistry	46			26		44		53	.49		.28	.27	1
#Reynolds and Hansvick (1978)	Psychology	47	14							.20				
Rodin and Rodin (1972)	Calculus	48		4	27			24	12	−.75	−.75	−.75	−.75	1
#Rubenstein and Mitchell (1970)	Psychology	49	15				45			.14				
Sheets et al. (1995).01	Microeconomics							25	58		.18	.15	.18	0
Sheets et al. (1995).02	Macroeconomics							26	63		−.14	−.25	−.14	0
!!Shmanske (1988)	Economics							27	17		.21			
Solomon et al. (1964)	American government	50					46		24	.46		.30	.19	1
Soper (1973)	Economics							28	14		−.17	−.17	−.17	0
Sorge and Kline (1973)	Mathematics	51							10	.39				
#Spencer and Dick (1965)		52								.88				
Sullivan and Skanes (1974).01	Science 115A	62			28	37		30	14	.42	.42	.51	.51	0
-Sullivan and Skanes (1974).02	Psychology-GTA	54			36	29			13	.01				
-Sullivan and Skanes (1974).03	Psychology-FT	55			35	30			27	.53				
Sullivan and Skanes (1974).04	Physics 1050	56			30	35		36	9	.57	.57	.57	.57	0
Sullivan and Skanes (1974).05	Math 1150	57			31	34		35	9	.33	.33	.33	.33	0
Sullivan and Skanes (1974).06	Match 1010	58			32	33		34	16	.34	.34	.34	.34	0
Sullivan and Skanes (1974).07	Match 100F	59			33	32		33	8	.48	.48	.48	.48	0
Sullivan and Skanes (1974).08	Chemistry 1000	60			34	31		32	6	.55	.55	.55	.55	0
Sullivan and Skanes (1974).09	Chemistry 100F	61	16		29	36		31	8	.08	.08	.08	.08	0
Sullivan and Skanes (1974).10	Biology 1010	53			37	28		38	14	.51	.51	.42	.42	0
Sullivan and Skanes (1974).11	Biology 100F	63				38		29	6	−.28	−.28	−.28	−.28	0
Sullivan and Skanes (1974).12	Psychology							37	40		.40	.40	.40	0
Turner and Thompson (1973).01	Beginning French	65					47		16	−.15		−.51	−.52	1
Turner and Thompson (1973).02	Beginning French	64					48		24	.08		−.41	−.38	1
Weinberg and Hashimoto (2007).01	Microeconomics							39	190		−.02	.04	.04	1
Weinberg and Hashimoto (2007).02	Macroeconomics							40	119		−.05	−.26	−.26	1
Weinberg and Hashimoto (2007).03	Microeconomics (intermed)							41	85		−.17	−.09	−.09	1
#Wherry (1952)	Psychology	66	17		38					.16				
Whitely and Doyle (1979).01	Math (intro)	67			40	39			5	.80		.80	.80	1
Whitely and Doyle (1979).02	Math (intro)	68			41	40			11	−.11		−.11	−.11	1
Wiviott and Pollard (1974)	Educational Psychology								6			−.04	.00	0
Yunker and Yunker (2003)	Accounting (intro)							42	46		−.11	.19	.19	0

Notes. Co81 = Cohen (1981); Co82 = Cohen (1982); Do82 = Dowell and Neal (1982); Co83 = Cohen (1983); Mc84 = McCallum (1984); Fe89 = Feldman (1989); Cl09 = Clayson (2009); *n* = number of sections; CIS = Current meta-analysis instructor only SET ratings only; CAS = Current meta-analysis average of all SET ratings; Adj = 1 if SET/learning correlations were adjusted for prior learning/ability and 0 if they were not. !! = did not meet inclusion criteria (see text) # = inaccessible – published as internal reports or presented in conferences only.

findings. He did not offer this possibility in his discussion of moderating factors and he did not discuss it elsewhere.

Our main objective was to re-examine the evidence for the SET/learning correlations and for the extant claims that SETs are valid measures of professors' teaching effectiveness rather than measures of student satisfaction. First, we re-analyzed the evidence provided by the previous meta-analyses of SET/learning relationships. For this purpose, we re-analyzed the data in Cohen's (1981) highly cited meta-analysis, Feldman's (1989) highly cited meta-analysis, and the most recent meta-analysis done by Clayson (2009). We did not re-analyze the data from other meta-analyses because they (a) were subsets of previous meta-analyses and/or not comprehensive (Cohen, 1982, 1983; Dowell & Neal, 1982; McCallum, 1984); (b) did not specify which studies they were based on (Cohen, 1982); and/or (c) were based on substantially incorrect data (McCallum, 1984). We took the data as presented by the authors, examined them for accuracy and re-analyzed them using both fixed and random effect meta-analyses. Critically, we examined each meta-analysis for the presence of small study effects using scatterplots, funnel plots, and regression tests, and then estimated SET/learning correlations by adjusting for the small study effects using several methods including the trim-and-fill estimate (Duval & Tweedie, 2000), the cumulative meta-analysis starting with the largest sample study and adding the next smaller study on each successive step (Rothstein, Sutton, & Borenstein, 2006), the estimate based on all studies with the sample equal or greater to 30 (NGT30), the estimate based on the top 10% (TOP10) of the most precise studies (Stanley & Doucouliagos, 2014), and the regression based estimate using limit meta analysis (Rücker, Schwarzer, Carpenter, Binder, & Schumacher, 2011). Second, because of the serious shortcomings of these prior meta-analyses, we conducted a comprehensive meta-analysis of SET-learning relationships from the ground up, starting with our own search for multisection studies.

1. Review and re-analysis of Cohen's (1981), Feldman's (1989), and Clayson's (2009) meta-analyses

1.1. Cohen (1981) meta-analysis

Cohen's (1981) meta-analysis was based on 68 multisection studies extracted from 41 published articles and other reports. Cohen found overall instructor and overall course ratings correlated with student achievement with $r = .43$ and $r = .47$, respectively. In addition, Cohen also analyzed correlations between several aspects of SET ratings (skill, rapport, structure, difficulty, interaction, feedback, evaluation, student progress) and

achievement. However, only two of the SET aspects were significantly correlated with achievement (skill: $r = .50$; structure: $r = .47$).

The review of Cohen's (1981) meta-analysis reveals numerous fundamental problems. First, the meta-analysis is lacking even the most essential details. To illustrate, Cohen did not report necessary details of his search for primary studies, did not report characteristics of primary studies, and did not even report effect size and study size (i.e., number of sections) for each individual multisection study. Second, the meta-analysis also revealed some impossibly high correlations ($rs > .90$) between SETs and learning, so called "voodoo" correlations (Vul, Harris, Winkielman, & Pashler, 2009; see, for example, Cohen's Fig. 2). Cohen (1981) noted that some of the reviewers of his article were concerned that the SET/learning correlations varied by the sample size but concluded that the number of sections did not have appreciable effect on the SET/learning correlations. Third, although Cohen (1981) was aware that some of the multisection studies were based on as few as five sections and that more than one third of his multisection studies had ten or fewer sections, Cohen disregarded the size of the individual multisection studies when he calculated the average correlation between SET and learning. Specifically, he combined the multisection study effect sizes by transforming *r*s to Fisher's *Z* scores, calculating average Fisher's *Z* score across all studies without weighing *Z*s by each study size, and transformed average Fisher *Z* scores back to *r*. Thus, Cohen gave equal weight to each primary study regardless of how many sections it was based on.

In an attempt to track down how Cohen (1981) arrived at his strong conclusion about SET/learning relationships, we obtained a microfilm of Cohen's (1980) PhD dissertation, the basis of Cohen's (1981) article. Although Appendix C in Cohen's (1980) dissertation contained the effect sizes for each primary study, it did not include other details about primary studies such as the number of sections included in each of them. Accordingly, to re-analyze Cohen's (1981) data, we located the articles and reports included in Cohen's meta-analysis and extracted the sample sizes for each multisection study reported in Appendix C in Cohen's (1980) dissertation.

Table 2 shows the list of multisection studies included in previous as well as in the current meta-analysis. The studies are identified by the first author of the article, followed by year of publication, and by the multisection study number within each article. The numbers in the "Co81" column indicate that the specific multisection study was included in the Cohen (1981) sample and the number itself corresponds to the number given to each multisection study by Cohen (1980) in Appendix C, column 2 (p. 99–100) in his PhD dissertation. The multisection studies preceded by "!!" do not meet Cohen's own inclusion criteria and were excluded from our re-analysis: Borg and Hamilton (1956) and Morsh et al. (1956) were not conducted using college/university classes; Hoffman (1978).02 did not report any SET/achievement correlations as Cohen inappropriately used correlations between SET and students' own perceptions of their learning (Feldman, 1989); Sorge and Kline (1973) confounded SET/learning correlations with experimental manipulation administered to half of the 10 sections.

Fig. 2, top left panel, shows that the magnitude of the SET/learning correlations as a function of the multisection study size. The figure includes the correlation (linear) between the SET/learning correlations and study size. Moreover, the SET/learning correlations plotted above or below the curved lines are statistically significant. The figure indicates that (1) the number of sections included in multisection studies was generally small with the number of multisection studies based on as few as five sections, (2) many impossibly high correlations ($r > .90$) were obtained in multisection studies with a small number of sections, (3) the majority of reported rating/achievement correlations were not statistically significant, and (4) the magnitude of SET/achievement correlations decreased for larger sized studies in the expected, non-linear fashion. Fig. 2, right panel, shows a cumulative meta-analysis, with the largest study entered first and the next largest study entered on each subsequent step. The meta-analysis shows that the SET/learning correlation estimated using only the studies with 30 or more sections (NGT30) is .27 and the magnitude of the correlation increases as the smaller studies are added into subsequent meta-analyses. Similarly, the SET/

Table 3
Previous meta-analyses of SET/learning correlations: Original analyses and current re-analyses.

	Original *r* (95% C.I.)	Original *n*	*r*	*n*	FAT *p*	RE (95% C.I.)	TF (#)	NGT30	TOP10	LMT (95% C.I.)
Cohen (1980)										
Overall Instructor	.43 (.21,.61)	67	.43	62	.055	.38 (.30,.44)	.33 (10)	.27	.27	.27 (.10,.42)
Overall Course	.47 (.09,.73)	22	.46	21	.013	.35 (.22,.47)	.29 (3)	.25	.27	.11 (-.15,.35)
Skill	.50 (.23,.70)	40	.54	37	< .001	.43 (.34,.51)	.35 (10)	.27	.29	.17 (-.03,.37)
Structure	.47 (.11,.72)	27	.47	27	<.001	.37 (.24,.50)	.19 (11)	.03	−.03	.06 (-.21,.33)
Feldman (1989)										
Preparation (Dim. 5)	.57	28	.55	26	<.001	.43 (.27,.56)	.23 (10)	.11	.02	.05 (-.21,.31)
Clarity (Dim. 6)	.56	32	.47	27	<.001	.38 (.27,.47)	.28 (9)	.16	.06	.11 (-.15,.36)
Perc. Outcome (Dim. 12)	.46	17	.43	16	.004	.30 (.13,.46)	.18 (5)	.07	.07	.00 (-.33,.33)
Stimulation (Dim. 1)	.38	19	.38	16	.077	.34 (.21,.45)	.30 (3)	.07	.10	.06 (-.40,.50)
Clayson (2009)										
Overall Instructor	.33	42	.33	.40	<.001	.29 (.16,.41)	.08 (15)	−.02	−.07	.06 (-.06,.17)

Note. FAT = Funnel Asymmetry Test via linear regression; RE = Random Effect *r*; TF = Trim and Fill *r* with # of imputed values in parentheses; NGT30 = *r* based on all studies with 30 or more sections; TOP10 = *r* based on top 10% of most precise/largest studies; LMT = adjusted *r* based on limit meta-analysis.

achievement correlation estimated using the top 10% of the sample (TOP10), or the six largest studies, is .27. Finally, Fig. 2, bottom left panel, shows the funnel plot with the estimated effect size adjusted for small-study effect using a regression based limit meta-analysis (Rücker et al., 2011). In this plot, the gray curve starts from a biased effect estimate for small size studies (bottom) to the adjusted estimate for a study with infinite precision (top). The adjusted $r = .27$ is printed in the top left corner of the panel.

Table 3 shows SET/learning correlations as reported by Cohen (1981) and the results of our re-analyses of Cohen's data for overall instructor rating and for overall course rating, as well as for two specific dimensions of the SET ratings that Cohen reported as significantly related to learning. First, the table shows SET/learning correlations re-calculated following Cohen's method (i.e., unweighted Fisher's Z transformed estimates that give equal weight to all studies regardless of the study size) but with several studies not meeting Cohen's own criteria excluded. The SET/learning average unweighted correlations calculated by Cohen and by us are either identical or nearly identical. Second, the table shows the summaries of our re-analyses of Cohen's (1981) data set following our analytical strategy outlined above. The regression test for funnel plot asymmetry was significant in all cases, suggesting the presence of small study bias. The random effects correlations weighted by study size (but not corrected for small study effects) were all smaller than the unweighted correlations reported by Cohen (1981). The trim-and-fill analyses resulted in adding three to 11 studies and still smaller estimates of the SET/learning correlations. The SET/learning correlations estimated using the NGT30, TOP10, and adjusted correlations using limit meta-analysis were still smaller. The limit meta-analysis estimates were substantially smaller for all SET/learning correlations and remained significant only for the overall instructor rating, that is, one out of four correlations reported as significant by Cohen (1981).

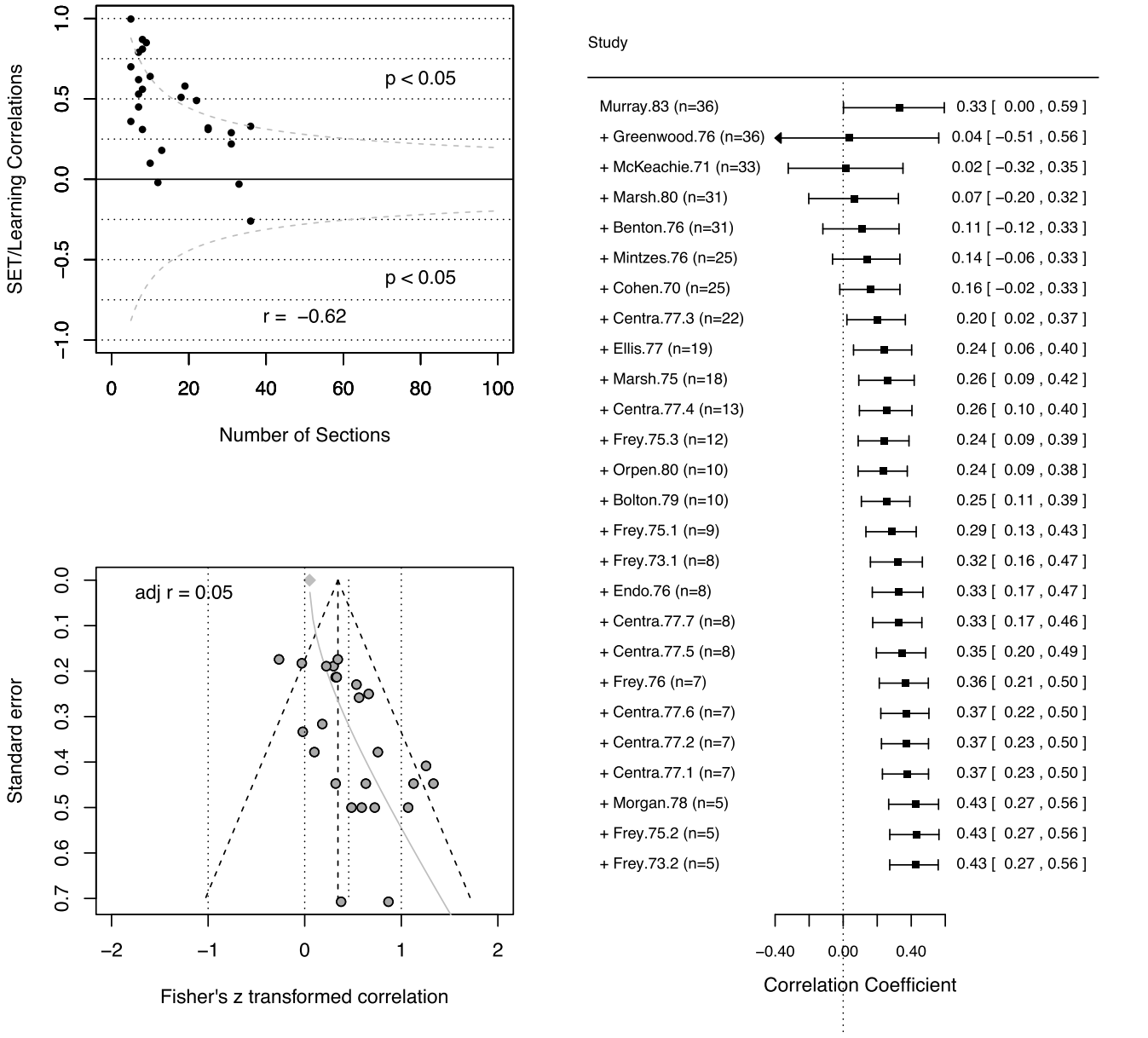

Fig. 3. Re-analysis of Feldman's (1989) multisection data set for Feldman's Dimension 5: Preparation and Organization. The top left panel shows the small size study effects. The right panel shows cumulative meta-analysis showing that the meta-analysis of large sized studies reveal small SET/learning correlation and that addition of smaller size studies increases the estimated SET/learning correlation. The bottom left panel shows the result of limit meta-analysis taking into account small size study effects, including adjusted $r = .05$.

In summary, Cohen's (1981) conclusion that SET/learning correlations are substantial and that SET ratings explain 18–25% of variability in learning measures is not supported by our re-analyses of Cohen's own data. The re-analyses indicate that SET ratings explain at best 10% of variance in learning measures. The inflated SET/learning correlations reported by Cohen appear to be an artifact of small study effects, most likely arising from publication bias.

1.2. Feldman (1989) meta-analysis

Feldman (1989) pointed out that Cohen's (1981) examination of the SET/learning correlations for specific SET dimensions was limited to only eight dimensions and that many (hundreds) of the reported SET/learning correlations in individual multisection studies were never reported nor analyzed by Cohen. Accordingly, Feldman's (1989) goal was to extend Cohen's (1981) meta-analysis by examining a correlation between 31 SET dimensions and learning, a much broader range of the dimensions than that reported by Cohen (1981). Feldman (1989) reported that some SET dimensions had moderate-to-strong correlations with measures of learning, with the four strongest SET/learning correlations ranging from .36 to .57.

The review of Feldman's (1989) meta-analysis reveals a nearly identical set of problems as those plaguing Cohen's (1981) meta-analysis. First, as noted above, Feldman (1989) did not conduct his own search for relevant articles/reports but simply relied on articles identified by Cohen (Abrami, Cohen, & d'Apollonia, 1988). In total, Feldman identified 46 articles cited in Cohen's work. However, because some of these articles did not include any SET/learning correlations for specific dimensions, Feldman's meta-analysis is based on 32 articles/reports that included 48 multisection studies with relevant dimension specific SET/learning correlations (Feldman, 1989). Second, Feldman (1989) did not consider the possibility that his results may be an artifact of small study effects. Third, Feldman also disregarded the size of individual multisection studies when he calculated average SET/learning correlations.

In Table 2, column "Fe89" identifies multisection studies included in Feldman's (1989) meta-analysis. The numbers refer to specific multisection studies listed in Feldman's Appendix B when the studies are numbered by the order of their listing in the appendix. Feldman (1989) noticed that Cohen (1981) used incorrect correlations for Hoffman (1978).02, and given that Cohen (1981) merely stated that the relevant correlations were not significant without actually providing them, Feldman replaced them with zeros (p. 642). We decided to exclude Hoffman (1978).02 from the re-analyses as the vast majority of reported SET/learning correlations were not statistically significant either.

Fig. 3, top left panel, shows the magnitude of the SET/learning correlations for Preparation & Organization (Feldman's dimension No. 5), the dimension most strongly correlated with learning in Feldman's analysis, as a function of the multisection study size. Similarly to the re-analyses of Cohen's (1981) data, the figure indicates that (1) a number of sections within each multisection study was generally small, (2) impossibly high correlations ($r > .90$) were obtained in multisection studies with small numbers of sections, (3) the majority of reported SET/learning correlations were not statistically significant, and (4) the magnitude of SET/learning correlations decreased for larger sized studies in an expected, non-linear fashion. Fig. 3, right panel, shows a cumulative meta-analysis, starting with the largest study and adding smaller studies in each successive step. The meta-analysis shows that the SET/learning correlation estimated using only the studies with 30 or more sections (NGT30) is .11 and the magnitude of the correlation increases as the smaller studies are added into subsequent meta-analysis. Similarly, the SET/learning correlation estimated using the TOP10 (i.e., the 3 largest studies) is .02. Finally, Fig. 3, bottom left panel, shows the result of the regression based limit meta-analysis, including the adjusted $r = .05$. We repeated these analyses for each of the next three dimensions with the strongest SET/learning correlations in Feldman's (1989) analyses: Clarity and Understandableness (Feldman's dimension No. 6), Perceived Outcome (Feldman's dimension No. 12), and Teacher's Stimulation of Interest in the Course and Its Subject Matter (Feldman's dimension No. 16). Similarly to Feldman's dimension No. 5, these re-analyses of the data showed that the SET/learning correlations for these dimensions were not statistically significant and negligible when the meta-analyses account for the small study effects (see Table 3).

Table 3 shows SET/learning correlations reported by Feldman (1989) for the four SET dimensions most strongly correlated with learning and the results of our re-analyses of Feldman's data. First, the table shows SET/learning correlations re-calculated following Feldman's method (i.e., unweighted Fisher's Z transformed estimates) are nearly identical or very similar to the correlations reported by Feldman. Second, the table shows that the regression test for funnel plot asymmetry was significant in all cases, suggesting the presence of small sample bias. The random effects correlations weighted by study size (but not corrected for small study effects) were all smaller than the unweighted correlations reported by Feldman (1989). The trim-and-fill analyses resulted in adding three to 10 studies and still smaller estimates of the SET/learning correlations. The SET/learning correlations estimated using the NGT30, TOP10, and adjusted correlations using limit meta-analysis were still smaller, negligible, and not statistically significant.

In summary, our re-analyses of Feldman's (1989) data indicate that Feldman's findings were an artifact of small study size effects and that his conclusion that the specific SET dimensions explain up to 33% of variance in learning is unwarranted. The re-analyses indicate that the specific SET dimensions do not significantly correlate with learning.

1.3. Clayson (2009) meta-analysis

The most recent meta-analysis of SET/learning correlations appeared nearly 30 years after Cohen (1981) and 20 years after Feldman (1989). Clayson (2009) used "generally, the same criteria as in the historical meta-analysis by Cohen (1981)" and located 17 articles with 42 multisection studies. Notably, Clayson mixed in Cohen's (1981) meta-analysis as if it were a multisection study reporting SET/learning correlation of .41 and having 35 sections. Clayson explained: "Cohen's (1981) meta-analysis contained 67 studies; those utilized in other places in this report were mathematically removed from Cohen's data. His average r with 67 cases was .43." (p.22). Using raw averaged correlation coefficients rather than Fisher's Z transformed coefficients, Clayson reported that the unweighted average SET/learning correlation was .33 whereas the weighted average correlation was only .13 and not statistically significant. Importantly, Clayson (2009) noticed that (1) the correlation between the SET/learning correlations (using Fisher's Z) and the sample size was negative (-.37), and that (2) the correlation between the SET/learning correlations (using Fisher's Z) and the publication's age was positive (.48), that is, the earlier studies showed larger SET/learning correlations than the later studies.

Unfortunately, Clayson's (2009) meta-analysis suffers from many of the same problems as the previous meta-analyses, as well as other problems that render Clayson's results largely uninterpretable. First, Clayson's (2009) description of his search for

relevant studies is too vague to be replicable. For example, Clayson did not specify in full which databases he searched and for which specific terms. Moreover, he reported that only "17 articles were found that contained 42 studies, including 1115 sections" (p. 21). Clearly, the search was not adequate as the number of articles included in the previous meta-analyses exceeds 40 articles. Cohen (1981) alone found 41 articles with 68 multisection studies. Second, as noted above, Clayson used Cohen's (1981) meta-analysis as one of his multisection studies with SET/learning correlation of .41 and 35 "sections." We cannot think of any reason justifying mixing the meta-analysis estimated *r* with multisection studies' *r* to conduct another meta-analysis of multisection *r*s. While such an approach may allow calculations of average raw correlations, it does not allow the calculation of average weighted correlations across the multisection studies because the weighted *r* is dependent on the size of each multisection study and because the 35 studies remaining in Cohen's (1981) meta-analysis represent hundreds of sections rather than 35 sections. As a result, Clayson's weighted average *r* is incorrect, artificially reduced, and ought not to be interpreted as it is meaningless. Finally, despite the fact that Clayson noticed the moderately strong correlations between SET/learning correlations and study size, Clayson did not investigate further and did not attempt to estimate the SET/learning correlations accounting for the small size effects.

Table 3, column "Cl09" identifies multisection studies, including Cohen's (1981) meta-analysis masquerading as a multisection study, included in Clayson's (2009) meta-analysis. The numbers refer to specific multisection studies listed in Clayson's Table 1 (p. 22) when the studies are numbered by the order of their listing. Two of these studies were excluded from the meta-analysis because they did not conform to Clayson's inclusion criteria: Cohen (1981) is not a multisection study and Shmanske (1988) did not use a common test to assess learning.

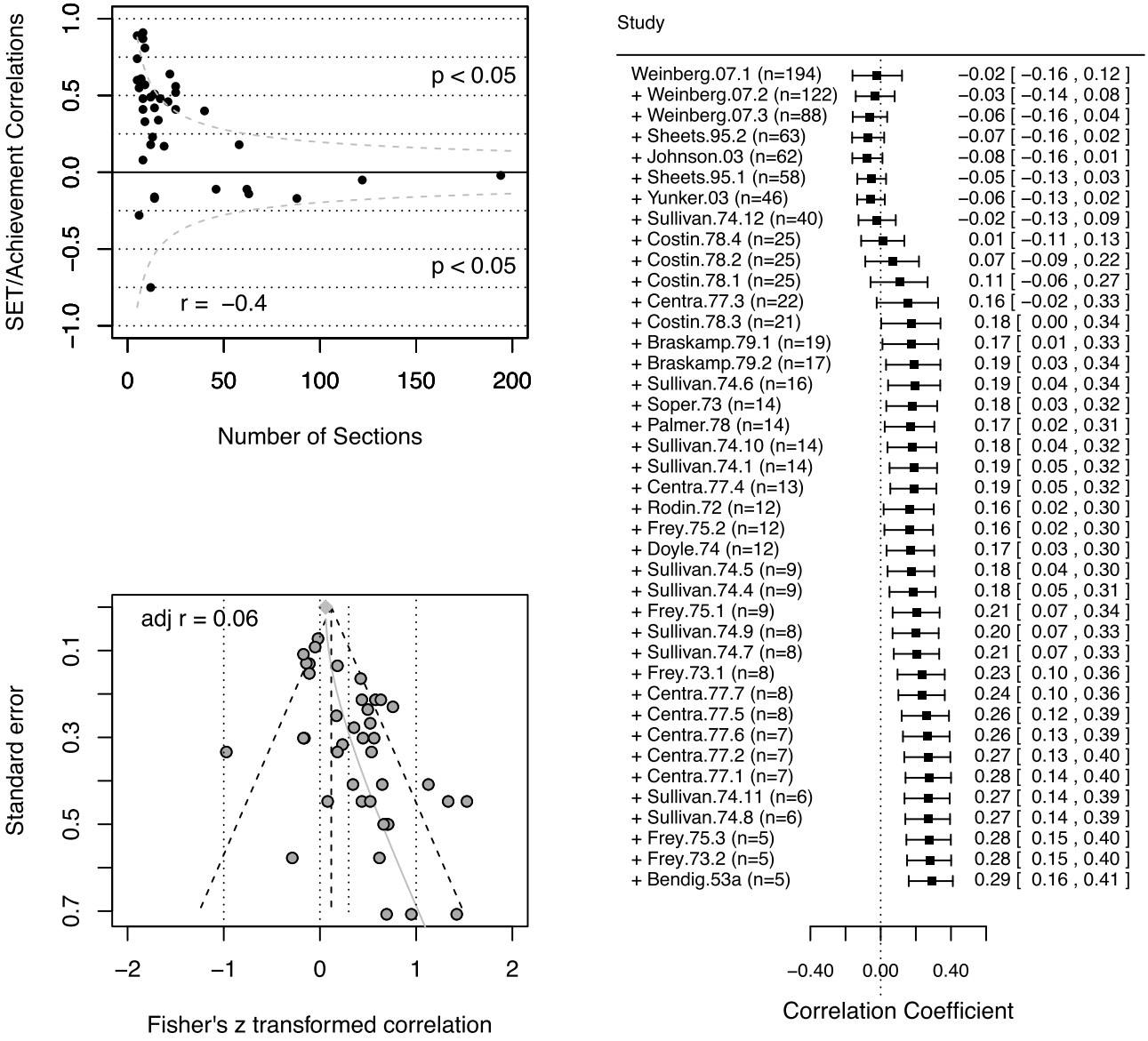

Fig. 4. Re-analysis of Clayson's (2009) multisection data set. The top left panel shows the small size study effects. The right panel shows cumulative meta-analysis showing that the meta-analysis of large sized studies reveals minimal and negative SET/learning correlation and that addition of smaller size studies increases the estimated SET/learning correlation. The bottom left panel shows the result of limit meta-analysis taking into account small size study effects, including adjusted *r* = .06.

Fig. 4 reveals the familiar pattern of findings. Fig. 4, top left panel, shows the magnitude of the SET/achievement correlations as a function of the multisection study size. The figure indicates that (1) a number of sections within each multisection study was generally small, (2) impossibly high correlations ($r > .90$) were observed in studies with a small number of sections, (3) the majority of reported SET/learning correlations were not statistically significant, and (4) the magnitude of SET/learning correlations decreased for larger sized studies in an expected, non-linear fashion. Importantly, the figure also shows that the new large sample studies show non-significant SET/learning correlations. Fig. 4, right panel, shows the cumulative meta-analysis, starting with the largest study and adding smaller studies in each successive step. The meta-analysis shows that the SET/learning correlation estimated using only the studies with 30 or more sections (NGT30) is −.02 and the magnitude of the correlation increases as the smaller studies are added into the subsequent meta-analyses. Similarly, the SET/learning correlation estimated using the TOP10 (the three largest studies) is slightly negative, −.07. Finally, Fig. 4, bottom left panel, shows the result of the regression based limit meta-analysis, including the adjusted $r = .06$. The summary of our re-analyses of Clayson's data are also in Table 3.

Accordingly, the re-analyses of Clayson's (2009) data reveal the same pattern of findings: the estimated correlations are smaller once the small study size effects are taken into account.

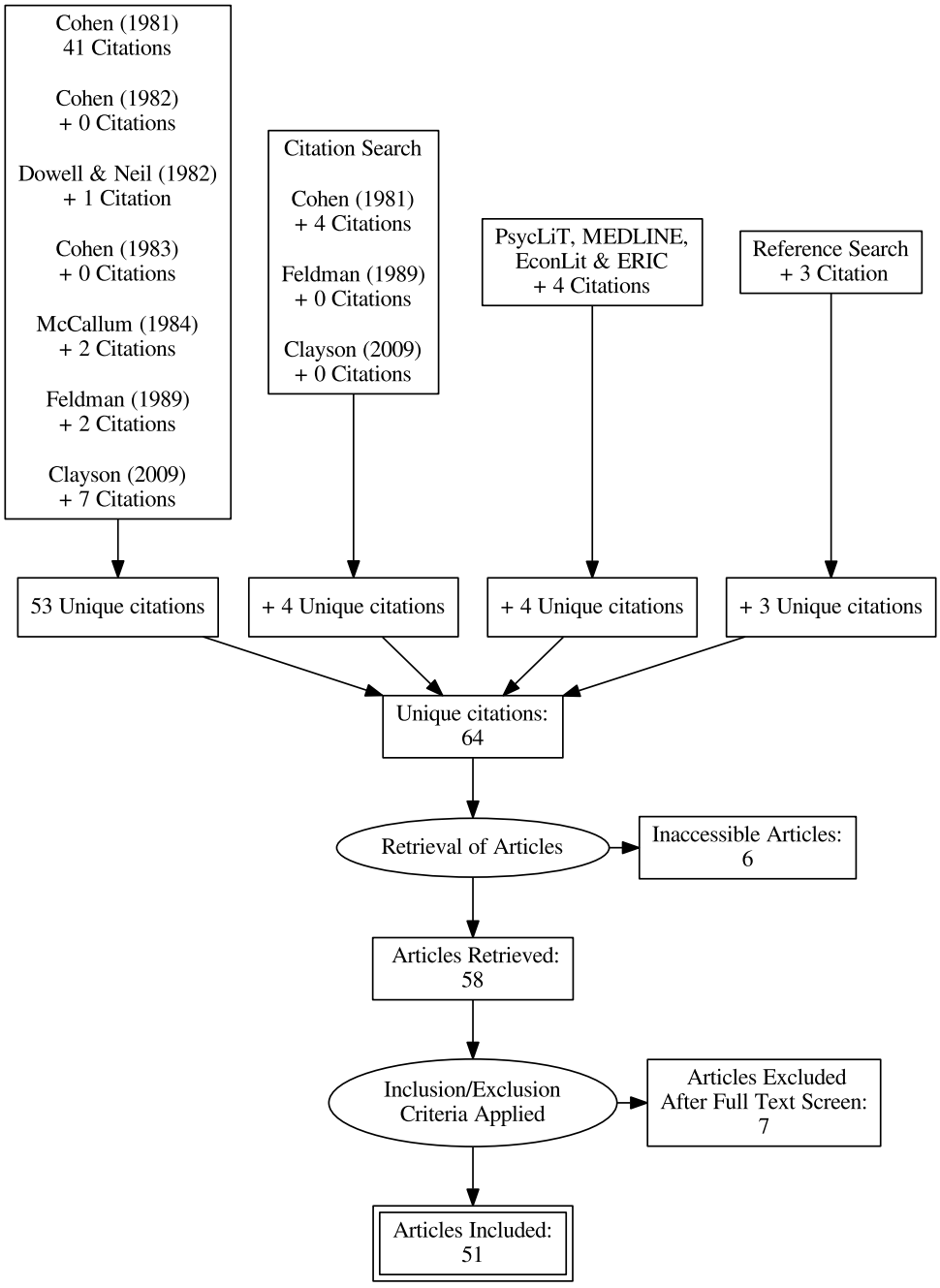

Fig. 5. Flowchart of the search strategy and selection of studies included in the review and meta-analysis.

The re-analyses indicate that the SET ratings do not significantly correlate with measures of learning. More critically, Clayson's meta-analysis is not comprehensive, included only a relatively small proportion of relevant multisection studies, and, oddly, included Cohen's meta-analysis as if it was a multisection study. As a result, Clayson's findings are largely uninterpretable and his weighted correlation estimate of SET/learning correlations is meaningless.

1.4. Summary

The review reveals that the previous meta-analyses suffer from numerous problems related to locating studies and none of the previous meta-analyses is replicable for that reason alone. Using the same inclusion criteria and explicitly searching for multisection studies in previous meta-analyses, Clayson (2009) should have found more articles and more multisection studies than Cohen (1981). More importantly, our review and re-analyses of the previous meta-analyses also indicates that the moderate SET/learning correlations reported in the previous meta-analyses are an artifact of small study size effects. The scatterplots and funnel plots of the SET/learning correlations as a function of study size reveal obvious small study size effects and the presence of these effects was confirmed by objective tests. Critically, when the SET/learning correlations are re-analyzed taking into account the small study size effects, the estimated SET/learning correlations drop to near zero for nearly all of the SET/learning correlations reported in the previous meta-analyses. Finally, the re-analyses of the previous data also indicate the presence of outliers (e.g., Rodin & Rodin, 1972) that previous meta-analyses did not mention nor consider whether or not they should be removed.

2. An up-to-date meta-analysis of SET/learning correlations reported by multisection studies

Given the shortcomings of prior meta-analyses of SET/learning correlations reported in multisection studies, we conducted a new comprehensive meta-analysis of the SET/learning correlations from the ground up. The new meta-analysis had several major aims. The first aim was to expand the set of multisection studies by including all multisection studies published to date. The second aim was to estimate SET/learning correlations in multisection studies while taking into account the presence of small study size effects. It is widely recognized that the issue of small study size bias often arising from a publication selection bias is a serious and common problem invalidating many meta-analyses (Rücker et al., 2011; Stanley & Doucouliagos, 2014). The third aim was to examine if SET/learning correlations were smaller in multisection studies that controlled versus did not control for prior learning/ability. The fourth aim was to examine SET/learning correlations for overall instructor ratings used in the previous meta-analyses of SET/learning relationships as well as SET/learning correlations for an *average* of SET/learning correlations reported in each multisection study. Multisection studies often report many more SET/learning correlations (i.e., SET/learning correlations for several items or SET factors) in addition to or instead of SET/learning correlations for overall instructor rating used in the previous meta-analyses. Moreover, although some universities and colleges use only overall instructor ratings to evaluate their professors, other universities and colleges use averages across all SET items or dimensions. Accordingly, we calculated the average SET/learning correlation reported by each multisection study and entered theses into separate meta-analyses. The fifth aim was to examine the sensitivity of the meta-analyses to extreme outliers visible in the previous meta-analyses.

3. Method

3.1. Studies included in meta-analysis

Fig. 5 shows the search for relevant studies which proceeded in several steps. First, the citations to articles with the multisection studies were collected from the previous meta-analyses (i.e., the meta-analyses listed in Table 1). Second, Web of Science Core Collection was searched for all articles citing Cohen (1981) and Feldman (1989). Because Clayson (2009) was not included in the Web of Science, Google Scholar was searched for all articles citing Clayson. Third, the PsycINFO, MEDLINE, EconLIT, and ERIC databases were searched from the earliest available date to the end of January 2016 for the following search terms: (a) TX "student* eval*" OR TX "student* rating*" OR TX "teach* effectiveness" OR TX "teach* performance", (b) TX "student* learning" OR TX "student* achievement" OR TX "academic achievement" or TX "student* performance," and (c) TX faculty OR TX professor* OR TX "teach* assistant*" OR TX instructor and the three searches were combined with AND. Fourth, the references in all relevant articles, book chapters, and theses, retrieved by any method, were examined for potentially relevant articles and the identified articles were hand searched for relevance. The relevant article may have reported on one or more multisection studies.

To be include in the meta-analysis, a study had to pass several inclusion criteria. First, the study had to report correlations or other measures of associations (e.g., regression, mean differences) between SET and learning/achievement in college or university settings. Second, each study had to involve multiple sections of the same rather than different courses. Third, the SET as well as the measures of learning had to be common for all sections within each study. Fourth, the learning measures had to be objective, assessing the actual learning rather than students' subjective perception of their learning. Fifth, the SET/learning correlations had to be calculated using section means rather than individual students' scores. And sixth, the study had to be written in English.

These criteria resulted in several exclusions. Two studies (Borg & Hamilton, 1956; Morsh, Burgess, & Smith, 1956) used in some of the previous meta-analyses were excluded because they did not examine SET/learning in college/university settings but in military training facilities, with the training completely dissimilar to typical college/university courses. Two studies (Gessner, 1973.01, Gessner, 1973.02) used in some of the previous meta-analyses were excluded because they did not involve multisection studies but rather different instructors teaching different modules of the same course (with modules confounded with instructors). Other studies were excluded for a variety of other reasons: Hoffman (1978).02 was excluded because it did not report necessary data to establish the SET/learning correlations; Sorge (1973) confounded SET/learning correlations with experimental manipulation; Johnson (2003) did not include a multisection study but a collection of sections from a variety of courses mixed together; Cohen (1981) was excluded because it did not include a multisection study but a meta-analysis of multisection studies; and Shmanske (1988) was excluded because the exam used differed among the sections.

Finally, six studies were excluded because they were inaccessible: Crooks and Smock (1974), Spencer and Dick (1965), and Wherry (1952) were internal reports, and Murray (1983), Reynolds and Hansvick (1978), and Spencer and Dick (1965) were conference presentations.

3.2. Recorded variables

For each multisection study, the recorded variables included: authors; year of publication; number of sections; SET/learning correlation; course name; course discipline; assignment of

students to sections (self-assigned, randomly assigned, other); prior ability/achievement controls (e.g., GPA, intelligence, pretest); SET measure; learning/achievement measure (e.g., final exam, final grade, proficiency exam); whether or not learning/achievement measure was common/same for all sections (different, not specified, same); learning/achievement measure objectivity (subjective, mixed, objective); instructor experience (e.g., graduate students, faculty, mix of graduate students and faculty, other, not specified); number of students in all sections; number of instructors; number of SET/learning correlations reported; publication venue; presence of conflict of interest (i.e., whether an author was involved in design or evaluation of SET used in the study; whether an author was associated with teaching/learning center, office responsible for evaluation of teaching, or commercial enterprise involved in selling SETs (e.g., ETS), or development of SET used in the study); publication venue (i.e., education, psychology, business/economics, or other journal). In addition, we also collected a number of measures of study quality including whether the study included means, SDs, ranges, reliabilities, and distributions for SET and learning/achievement measures; and whether a study included any scatterplots of SET/learning relationships.

3.3. Meta-analysis methodology

Some multisection studies reported only zero order SET/learning correlations, other multisection analyses reported SET/learning correlations adjusted for prior learning and/or ability, and still other multisection analyses reported both zero order and prior knowledge/ability adjusted SET/learning correlations. Given that

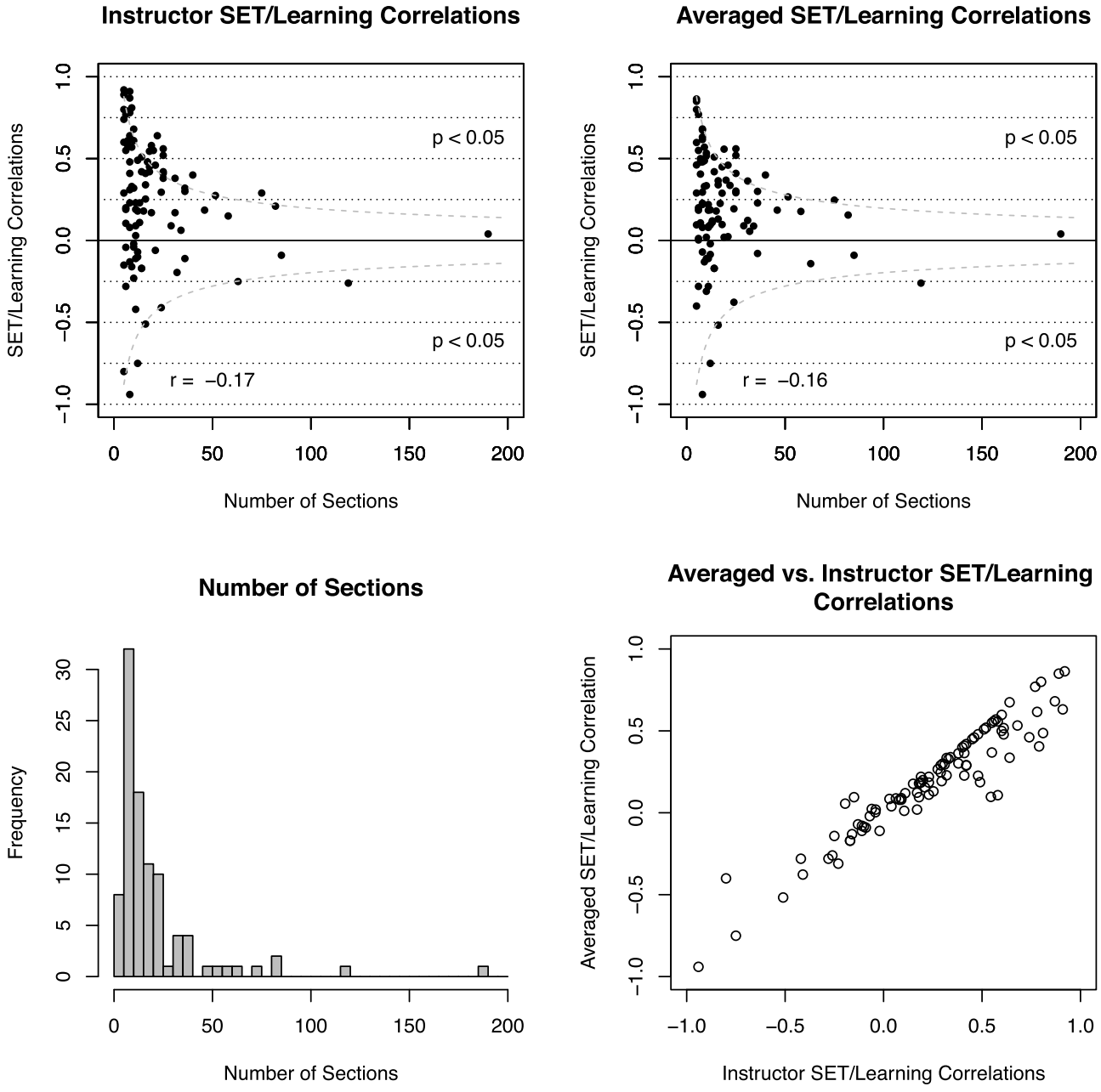

Fig. 6. The top left panel shows the relationship between Instructor SET/learning correlations and study size (number of sections). The top right panel shows the relationship between averaged SET/learning correlations and study size (number of sections). The bottom left panel shows the histogram of the number of sections. The bottom right panel shows the scatterplot between Averaged vs. Instructor SET/learning correlations ($r = .95$).

in nearly all multisection studies students were not randomly assigned to the sections, we followed the previous meta-analyses and used the SET/learning correlation adjusted for prior knowledge/ability and we used zero order correlations only if the prior knowledge/adjusted correlations were not available. Consistent with our aims, we tested whether the type of the best available SET/learning correlation – zero order or adjusted for prior knowledge/ability – moderates the SET/learning relationship. Finally, we conducted separate meta-analyses using only multisection studies that provided knowledge/ability adjusted correlations because even if the moderator test was not statistically significant, the prior knowledge/ability adjusted correlations are the better estimate of learning than zero order correlations.

Some multisection studies reported only one SET/learning correlation, typically between overall instructor SET rating and learning/achievement. Other multisection studies reported a number of SET/learning correlations, for example, one for each SET item. Accordingly, we analyzed the data two ways. First, in the first set of meta-analyses, for each multisection study, we used only one SET/learning correlation, that is, the one that best captured the correlation between overall instructor rating and learning/achievement. This approach follows Cohen (1981) as well as Clayson (2009). For the second set of meta-analyses, for each multisection study, we used averaged SET/learning correlations averaged across all SET/learning items. However, we never averaged across zero order and ability/prior achievement adjusted correlations.

We examined the data for the presence of outliers and small study effects using boxplots, scatterplots, funnel plots, and regression tests. Next, we estimated effect size using the random effect model (using restricted maximum-likelihood estimator or REML) but also provided fixed effects estimates for basic analyses and for comparison with prior meta-analyses. A random effect model allows for true effect size to vary from study to study, for example, the effect size may be a little higher for some academic disciplines than for other academic disciplines or it may be higher for studies conducted in colleges than for studies conducted in universities. In contrast, the fixed effect model assumes that all primary studies provide estimate of a single true effect size. Given variety of disciplines, institutions, SET measures, learning measures, etc. employed by primary studies, the key assumption of fixed effect model is unlikely to be true and the random effect model is more appropriate. We supported these analyses with forest plots. Next, we estimated SET/learning correlations adjusted for the small study effects using several basic as well as more sophisticated methods, including the trim-and-fill estimate, the cumulative meta-analysis starting with the largest sample study and adding the next smaller study on each successive step, the estimate based on all studies with the sample equal or greater to 30 (NGT30), the estimate based on the top 10% of the most precise studies (TOP10), and the regression based estimates using limit meta-analysis method. In general, based on a variety of simulation studies, when small study effects are present, the TOP10 and the regression based estimates using the limit meta-analysis method perform the best (Moreno et al., 2009; Stanley & Doucouliagos, 2014). Next, we also examined the sensitivity of meta-analyses to outliers. All reported analyses were conducted using R, and more specifically, using packages meta, metafor, and metasens.

4. Results

The 51 articles yielded 97 multisection studies. Table 2 shows overall instructor SET/Learning correlations (column labeled "CIS *r*") as well as averaged SET/Learning correlations (column labeled "CAS *r*") across all items/factors for each multisection study. Fig. 6 shows the relationship between Instructor SET/Learning correlations and study size (number of sections), the relationship between Averaged SET/Learning correlations and study size

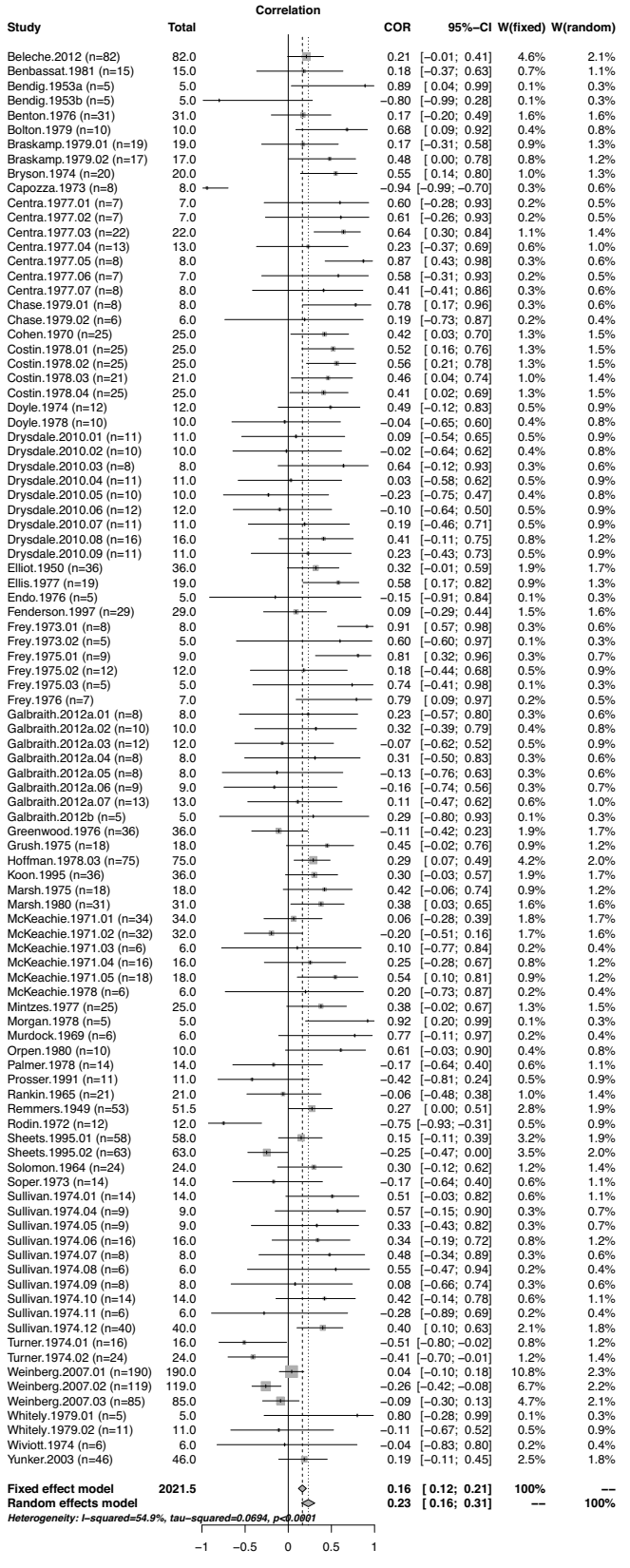

Fig. 7. Forest plot for Instructor SET/learning correlations. The plot includes the study identifier, number of sections, correlation, 95% C.I., and weights for each study as well as fixed effects and random effects estimates.

(number of sections), the histogram of the number of sections included in multisection studies, and the scatterplot between Averaged versus Instructor SET/Learning Correlations. The figure confirms the presence of small size study effects in both Instructor SET/Learning correlations and in Averaged SET/Learning correlations in this expanded data set. Moreover, the histogram of multisection study sizes confirms that the majority of studies were based on a small number of sections and many on fewer than 10 sections. Finally, the scatterplot between Averaged and Instructor SET/Learning correlations indicates that the two sets of correlations are very highly related, $r = .95$, but that the Averaged SET/Learning correlations tend to be somewhat smaller in absolute value than Instructor SET/Learning correlations.

4.1. Overall instructor SET/learning correlations

Fig. 7 shows the forest plot and both fixed and random effects model meta-analysis for SET/learning correlations using all SETs. The random effects model ($k = 97$) shows $r = .23$ with 95% CI = (.16, .31), with a moderate heterogeneity as measured by $I^2 = 54.9\%$, $Q(96) = 212.73$, $p < .001$. Moreover, the mixed effects moderator analysis showed that SET/learning correlations were substantially smaller for studies with adjustment for prior knowledge/ability,

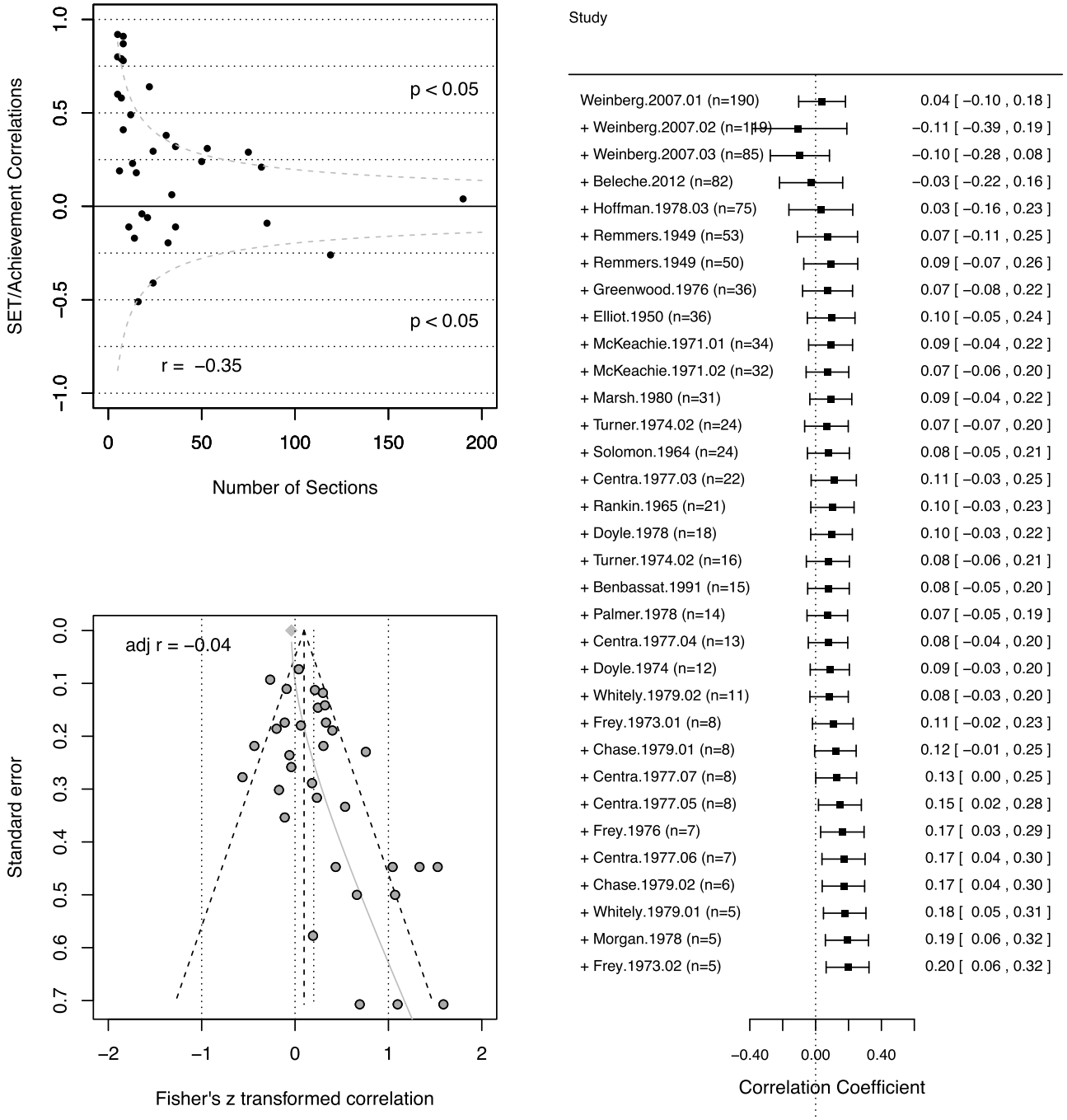

Fig. 8. Meta-analysis of Instructor SET/learning correlations using only correlations adjusted for prior learning/ability. The top left panel shows the small size study effects. The right panel shows cumulative meta-analysis showing that the meta-analysis of large sized studies (i.e, with > 30 sections) suggests minimal and non significant SET/learning correlation and that addition of smaller size studies increases the estimated SET/learning correlation. The bottom left panel shows the result of limit meta analysis taking into account small size study effects, including adjusted $r = -.04$.

$r = .12$ with 95% CI = (0,.24) than for studies without such adjustments, $r = .30$ with 95% CI = (.20, .38), $Q(1) = 5.21$, $p = .022$. However, this estimate does not take into an account the presence of the small study effects. Using all studies, the linear regression test of funnel plot asymmetry indicated asymmetry, p = .002. The estimates of SET/learning r adjusted for small study effects were: TF: .12 (with 22 filled in effects); NGT30: .10; Top10: .08; and limit meta-analysis adjusted $r = .12$ with 95% CI = (.03, .21) (Test of small-study effects: Q-$Q'(1) = 21.24$, $p < .001$; test of residual heterogeneity $Q(95) = 191.49$, $p < .001$).

We re-ran the above analyses but only for studies with prior knowledge/ability adjustments. The random effect model ($k = 34$) shows $r = .16$ with 95% CI = (-.02,.32), with moderate heterogeneity as measured by $I^2 = 72.2\%$, $Q(33) = 118.92$, $p < .001$. The linear regression test of funnel plot asymmetry was not significant, $p = .113$. The estimates of SET/learning r adjusted for small study effects were: TF: −.01 (with 8 filled in effects); NGT30: .08; Top10: −.03; and limit meta-analysis adjusted $r = -.06$ with 95% CI = (−.17, .07) (Test of small-study effects: Q-$Q'(1) = 9.10$, $p = .003$; test of residual heterogeneity $Q(32) = 109.82$, $p < .001$).

Finally, the two studies – Capozza (1973) ($n = 8$) and Rodin and Rodin (1972) ($n = 12$) – who were identified as univariate outliers in the preliminary analyses were also extreme outliers with studentized residuals below −3.0. Accordingly, we re-ran the above analyses with these two studies removed. With the two outliers removed, the random effect model ($k = 95$) shows $r = .25$ with 95% CI = (.18, .31), with lower heterogeneity $I^2 = 48.0\%$, $Q(95) = 182.85$, p < .001. Moreover, the mixed effects moderator analysis showed that SET/learning correlations were substantially smaller for studies with adjustment for prior knowledge/ability, $r = .17$ with 95% CI = (.05, .27) than for studies without such adjustments, $r = .30$ with 95% CI = (.21, .38), $Q(1) = 3.34$, $p = .068$. However, as noted above, this estimate does not take into an account the presence of the small study effects. Using all studies, the linear regression test of funnel plot asymmetry indicated asymmetry, $p < .001$. The estimates of SET/learning r adjusted for small study effects were: TF: .13 (with 24 filled in effects), NGT30: .10, Top10: .08, and limit meta-analysis adjusted $r = .11$ with 95% CI = (.02, .20) (Test of small-study effects: Q-$Q'(df = 1) = 29.09$, p < .001; test of residual heterogeneity $Q(93) = 153.63$, $p < .001$).

We re-ran the above analyses but only for studies with prior knowledge/ability adjustments. The random effect model ($k = 32$) shows $r = .20$ with 95% CI = (.06, .34), with moderate heterogeneity as measured by $I^2 = 66.2\%$, $Q(31) = 94.48$, $p < .001$. The linear regression test of funnel plot asymmetry was significant, $p = .006$. We recalculated the estimates of SET/learning r adjusted for small study effects: TF: .04 (with 10 filled in effects), NGT30: .08, Top10: −.03, and limit meta-analysis adjusted $r = -.05$ with 95% CI = (-.17, .07) (Test of small-study effects: Q-$Q'(1) = 20.41$, $p < .001$; test of residual heterogeneity $Q(30) = 72.07$, $p < .001$). Fig. 8, top left panel, shows the magnitude of the SET/learning correlations as a function of the multisection study size revealing the familiar small study size effects. The Fig. 8, right panel, shows the cumulative meta-analysis, starting with the largest study and adding smaller studies in each successive step; it indicates that the magnitude of the correlation increases as the smaller studies are added into subsequent meta-analysis. Finally, Fig. 8, bottom left panel, shows the result of the regression based limit meta-analysis, including the adjusted $r = -.04$.

4.2. Averaged SET/learning correlations

Fig. 9 shows the forest plot and both fixed and random effects model meta-analysis for SET/learning correlations using all SETs. The random effect model ($k = 97$) shows $r = .17$ with 95% CI = (.11, .23), with a low heterogeneity as measured by $I^2 = 34.1\%$, Q

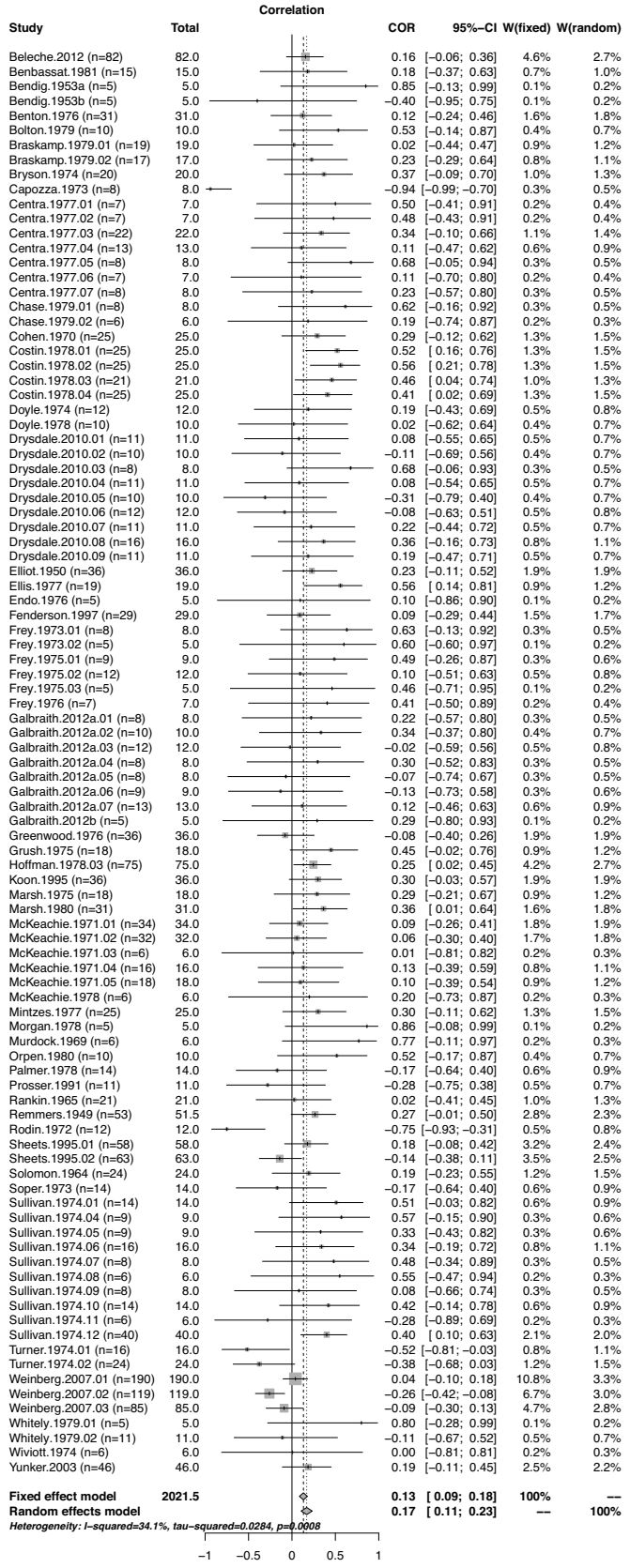

Fig. 9. Forest plot for Averaged SET/learning correlations, including fixed effect and random effects model estimates. The plot includes the study identifier, number of sections, correlation, 95% C.I., and weights for each study as well as fixed effects and random effects estimates.

(96) = 145.58, $p < .001$. Moreover, the mixed effects moderator analysis showed that SET/learning correlations were substantially smaller for studies with adjustment for prior knowledge/ability, $r = .05$ with 95% CI = (-.04, .14) than for studies without such adjustments, $r = .25$ with 95%CI = (.17, .33), $Q(1) = 10.46$, $p = .001$. However, this estimate does not take into account the presence of the small study effects. Using all studies, the linear regression test of funnel plot asymmetry indicated asymmetry, $p = .008$. The estimates of SET/learning r adjusted for small study effects were: TF: .10 (with 22 filled in effects); NGT30: .10; Top10: .08; and limit meta-analysis adjusted $r = .09$ with 95% CI = (0,.19) (Test of small-study effects: Q-$Q'(1) = 10.32$, $p = .001$; test of residual heterogeneity $Q(95) = 135.26$, $p = .004$).

We re-ran the above analyses but only for studies with prior knowledge/ability adjustments. The random effect model ($k = 34$) shows $r = .06$ with 95% CI = (-.06, .17), with moderate heterogeneity as measured by $I^2 = 57.0\%$, $Q(33) = 76.75$, $p < .001$. The linear regression test of funnel plot asymmetry was not significant, $p = .373$. However, in light of the overall asymmetry effects, we recalculated the estimates of SET/learning r adjusted for small study effects and provide them in Table 4 together with the summaries of the above analyses.

Finally, the two studies – Capozza (1973) ($n = 8$) and Rodin and Rodin (1972) ($n = 12$) – who were identified as univariate outliers in the preliminary analyses, were also extreme outliers with studentized residuals below −3.0. Accordingly, we re-ran the above analyses with these two studies removed. With the two outliers removed, the random effects model ($k = 95$) shows $r = .18$ with 95% CI = (.12, .24), with lower heterogeneity $I^2 = 19.6\%$, $Q(95) = 116.86$, $p = .055$. Moreover, the mixed effects moderator analysis showed that SET/learning correlations were substantially smaller for studies with adjustment for prior knowledge/ability, $r = .09$ with 95% CI = (-.01, .17) than for studies without such adjustments, $r = .25$ with 95% CI = (.17, .32). However, as noted above, this estimate does not take into account the presence of the small study effects. Using all studies, the linear regression test of funnel plot asymmetry indicated asymmetry, $p < .001$. The estimates of SET/learning r adjusted for small study effects were: TF: .10 (with 24 filled in effects), NGT30: .10, Top10: .08, and limit meta-analysis adjusted $r = .08$ with 95% CI = (−.01, .17) (Test of small-study effects: Q-$Q'(1) = 14.86$, $p < .001$; test of residual heterogeneity $Q(94) = 102.76$, $p = .252$).

We re-ran the above analyses but only for studies with prior knowledge/ability adjustments. The random effects model ($k = 32$) shows $r = .09$ with 95% CI = (-.01, .19), with moderate heterogeneity as measured by $I^2 = 39.8\%$. The linear regression test of funnel plot asymmetry was significant, $p = .027$. We recalculated the estimates of SET/learning r adjusted for small study effects: TF: .02 (with 9 filled in effects), NGT30: .08, Top10: −.04, and limit meta-analysis adjusted $r = -.03$ with 95% CI = (-.15,.09) (Test of small-study effects: Q-$Q'(1) = 7.81$, $p = .005$; test of residual heterogeneity $Q(30) = 43.64$, $p = .051$). Fig. 10, top left panel, shows the magnitude of the SET/learning correlations as a function of the multisection study size revealing the familiar small study size effects. Fig. 10, right panel, shows the cumulative meta-analysis indicating that the magnitude of the correlation increases as the smaller studies are added into subsequent meta-analysis. Finally, Fig. 10, bottom left panel, shows the result of the regression based limit meta-analysis, including the adjusted $r = -.03$.

5. Discussion

Our meta-analyses of SET/learning correlations reported in multisection studies reveals the following findings. First, multisection studies typically included a very limited number of sections, most employing 10 or fewer sections. Second, scatterplots of SET/learning correlations as a function of study size, funnel plots, and funnel asymmetry tests indicate presence of strong small study size effects. The small sized studies often reported impossibly high voodoo SET/learning correlations whereas the large sized studies reported small or no correlations. Third, when the analyses include both multisection studies with and without prior learning/ability controls, the estimated SET/learning correlations are very weak with SET ratings accounting for up to 1% of variance in learning/achievement measures. Fifth, when only those multisection studies that controlled for prior learning/achievement are included in the analyses, the SET/learning correlations are not significantly different from zero. Sixth, the above findings hold for both overall instructor SET ratings as well as for averages of all SET ratings reported by various multisection studies.

Table 4
Current meta-analyses of SET/learning correlations.

	r_{Zr}	r	RE (95% C.I.)	I^2	Q p	FAT p	TF (#)	NGT30	TOP10	LMT (95% C.I.)
Instructor SET only										
All data ($k = 97$)	.28	.24	.23 (.16,.31)	54.9	212.73 <.001	.002	.12 (22)	.10	.08	.12 (.03,.21)
Adjusted rs only ($k = 34$)	.25	.19	.16 (-.02,.32)	72.2	118.92 <.001	.113	−.01 (8)	.08	−.03	−.06 (-.17,.07)
Outliers removed ($k = 95$)	.31	.27	.25 (.18,.31)	48.0	182.85 <.001	<.001	.13 (24)	.10	.08	.11 (.02,.20)
Adjusted rs only ($k = 32$)	.34	.26	.20 (.06,.34)	66.2	94.48 <.001	.006	.04 (10)	.08	−.03	−.05 (-.17,.07)
Average of all SET										
All data ($k = 97$)	.22	.20	.17 (.11,.23)	34.1	145.58 <.001	.008	.10 (22)	.10	.08	.09 (0,.19)
Adjusted rs only ($k = 34$)	.13	.13	.06 (-.06,.17)	57.0	76.75 <.001	.373	0 (7)	.08	−.04	−.02 (-.14,.10)
Outliers removed ($k = 95$)	.25	.22	.18 (.12,.24)	19.6	116.86 .055	<.001	.10 (24)	.10	.08	.08 (-.01,.17)
Adjusted rs only ($k = 32$)	.22	.19	.09 (-.01,.19)	39.8	51.46 .012	.027	.02 (9)	.08	−.04	−.03 (-.15,.09)

Note. r_{Zr} = average unweighed Fisher's Z transformed r; r = average unweighted r; RE = Random Effect r. I^2 = heterogeneity index; Q = test of heterogeneity; FAT = Funnel Asymmetry Test via linear regression; TF = Trim and Fill r with # of imputed values in parentheses; NGT30 = r based on all studies with 30 or more sections; TOP10 = r based on top 10% of most precise/largest studies; LMT = adjusted r based on limit meta-analysis.

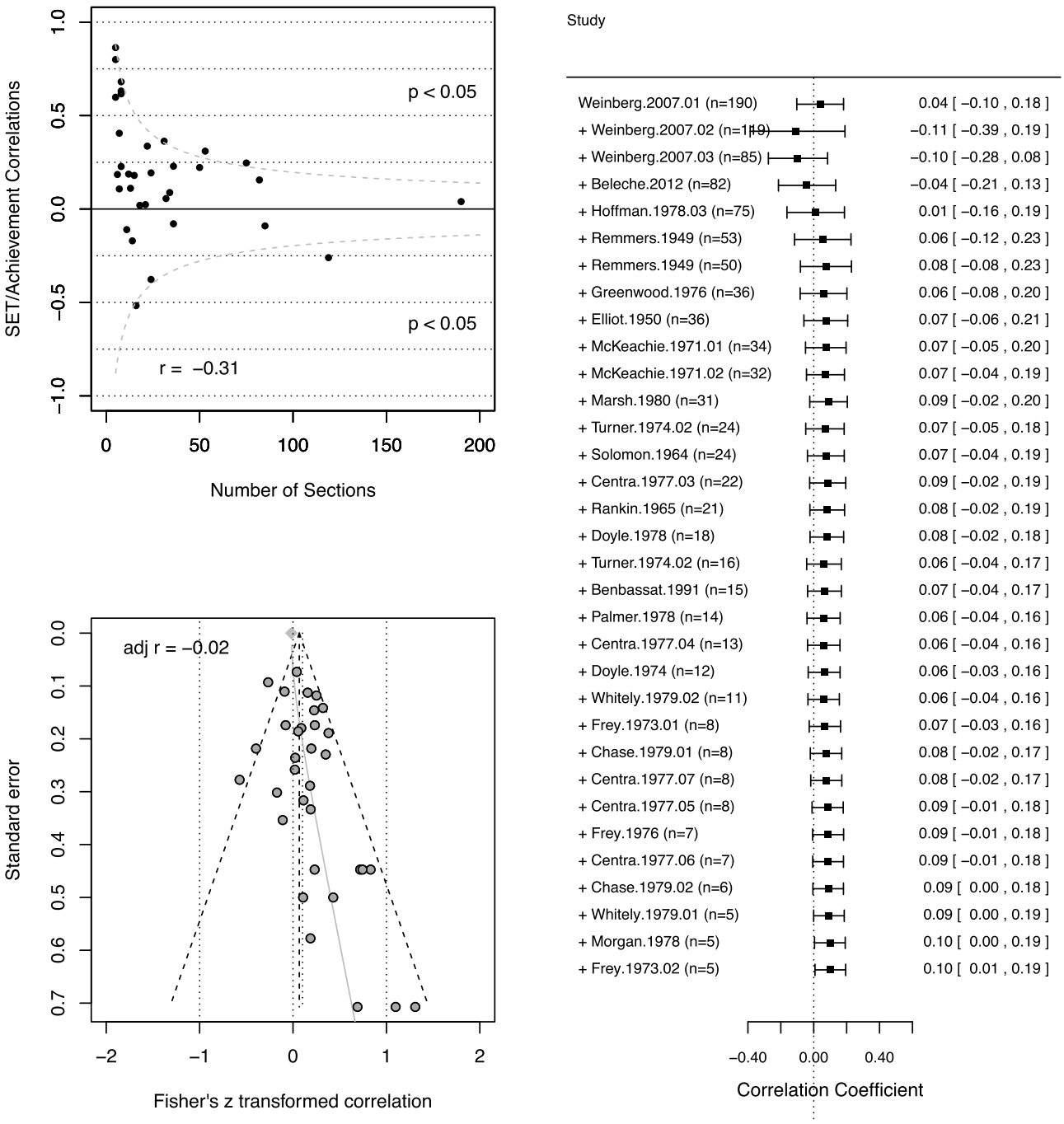

Fig. 10. Meta-analysis of Averaged SET/learning correlations using only correlations adjusted for prior learning/ability. The top left panel shows the small size study effects. The right panel shows cumulative meta-analysis showing that the meta-analysis of large sized studies reveals minimal and nonsignificant SET/learning correlation and that addition of smaller size studies increases estimated SET/learning correlation. The bottom left panel shows the result of limit meta-analysis taking into account small size study effects, including adjusted $r = -.02$.

Accordingly, the multisection studies do not support the claims that students learn more from more highly rated professors.

The review and re-analyses of the data sets reported in the previous meta-analyses indicate that at least Cohen (1981), Feldman (1989), and Clayson (2009) would have arrived to much the same conclusions if they (a) plotted the SET/learning correlations as a function of study size and (b) drew the most obvious conclusions from such scatterplots. Although not considered in the previous meta-analyses, the scatterplots of SET/learning correlations as a function of study size (i.e., number of sections within each study) indicate the presence of impossibly high voodoo correlations and strong small study size effects. When these small study size effects are taken into account, the estimated SET/learning correlations drop to near zero for nearly all of the SET/learning correlations reported in the previous meta-analyses. Specifically, the re-analyses of Cohen (1981) data do not support his claim that multisection studies provide "strong support for the validity of student ratings as measure of teaching effectiveness." Similarly, the re-analyses of Feldman (1989) data do not support the claim that various dimensions of SET ratings are strongly related to teaching effectiveness. Finally, the review and re-analysis of Clayson's (2009) data reveal the same small study size effects that invalidate some of Clayson's findings, and it highlights that Clayson's odd approach to meta-analysis – using Cohen's

(1981) meta-analysis as if it were a multisection study with 35 sections – invalidates most of his other analyses and findings including his estimate of weighted average SET/learning correlation.

In combination, our new up-to-date meta-analyses based on nearly 100 multisection studies, as well as our re-analyses of the previous meta-analyses make it clear that the previous reports of "moderate" and "substantial" SET/learning correlations were artifacts of small size study effects. The best evidence – the meta-analyses of SET/learning correlations when prior learning/ability are taken into account – indicates that the SET/learning correlation is zero. Contrary to a multitude of reviews, reports, as well as self-help books aimed at new professors (a few of them quoted above), the simple scatterplots as well as more sophisticated meta-analyses methods indicate that students do not learn more from professors who receive higher SET ratings.

It is astonishing that despite over 30 years of various reviews of multisection studies the reviewers have not noticed and not followed up on various red flags present in the previous meta-analyses, such as the impossibly high "voodoo" SET/learning correlations, small study sizes, and failures to report SET/learning correlations and study sizes for each multisection study. Even though Cohen himself acknowledged that the reviewers of his work were concerned about the small size study effects (Cohen, 1981), we were unable to locate a single scatterplot of these relationships in any of the prior meta-analyses nor in any of the major reviews. Similarly, Abrami et al. (1988) discovered evidence that there was something seriously wrong with the early meta-analyses but did not investigate further. For example, in their Table 3, Abrami et al. tabulated SET/learning correlations extracted from multisection studies by Cohen (1983), Dowell and Neal (1982), and McCallum (1984), observed "troublesome" disagreement between the correlations extracted by Cohen and McCallum, but did not pursue the matter further by, for example, locating the multisection studies and checking the extracted data for accuracy. If they did, they would have discovered that a large proportion of McCallum's data was simply incorrect, extracted from other irrelevant tables within the same multisection study articles. In turn, they would have had to conclude that none of McCallum's findings and conclusions were valid. Undoubtedly, any reviewer who actually looked at Cohen's (1981) article rather than relying on and trusting other reviewers was met with palpable absence of data and impossibility to assess the reasonableness of Cohen's findings and conclusions.

The entire notion that we could measure professors' teaching effectiveness by simple ways such as asking students to answer a few questions about their perceptions of their course experiences, instructors' knowledge, and the like seems unrealistic given well established findings from cognitive sciences such as strong associations between learning and individual differences including prior knowledge, intelligence, motivation, and interest. Thus, the individual differences in knowledge and intelligence are likely to influence how much students learn in the same course taught by the same professor. Similarly, individual differences in students' prior interest in a course are likely to influence how engaged they are, how hard they work and how much they learn. We (Uttl, White, & Morin, 2013) have recently shown that undergraduate students' interest in quantitative vs non-quantitative courses was very low with the students' mean interest in statistics courses nearly six standard deviations below their mean interest in non quantitative courses taught in the same psychology department. Fewer than 10 students out of 340 students responded that they were "very interested" in taking any of the statistical courses. In contrast, nearly half of the students were "very interested" in taking abnormal psychology course. Would we expect equally effective/competent professors teaching these courses, one populated with mostly disinterested students and one populated with mostly interested students, to receive the same SET ratings? Probably not. In fact, prior research indicates that prior interest in a course is one of the strongest predictor of SET ratings and that professors teaching quantitative courses receive lower SET ratings than professors teaching non quantitative courses. However, some speculate that quantitative courses receive lower SET ratings because professors teaching them may be less competent and less effective (Benton & Cashin, 2012).

It has been argued that SETs are responsible for grade inflation and work deflation in higher education by shifting the responsibility for students' learning and grades from students to professors. In response, Abrami and d'Apollonia (1999) opined:

> "... academic standards that are too high may be as detrimental to the learning of students as academic standards that are too low. The art and science of good teaching is finding the balance between what students might learn and what students are capable of learning. We believe that ratings help identify those instructors who do this well." (p. 520)

In this view, SETs are some sort of measurement instrument device enabling professors to find what students' perceive to be an appropriate workload and an appropriate amount to learn for specific grades, in short, an appropriate academic standard from students' perspectives. Professors who do this well, argue Abrami and d'Appolonia, will get high SETs. In contrast, professors who are either unable to do it well or do not do it because they believe that such student determined academic standards are detrimental to the students' themselves and/or to the society at large will get poor SETs. It follows that if the student determined standards are too far off from the standard necessary to pass the next course, attain a degree, or succeed in a new career after graduation, a professor is faced with a stark dilemma: teach to the SET and be promoted and tenured, or teach to prepare students for the next course, graduation and future careers, and be terminated.

In conclusion, two key findings emerged: (1) the findings reported in previous meta-analyses (Clayson, 2009; Cohen, 1981; Feldman, 1989) are an artifact of poor meta-analytic methods, and (2) students do not learn more from professors with higher SETs. The reported correlations between SET ratings and learning are completely consistent with randomly generating correlations from the population correlation with *rho* = 0 and applying publication selection bias. Despite more than 75 years of sustained effort, there is presently no evidence supporting the widespread belief that students learn more from professors who receive higher SET ratings. If anything, the latest large sample studies show that students who were taught by highly rated professors in prerequisites perform more poorly in follow up courses (Weinberg, Hashimoto, & Fleisher, 2009; Yunker & Yunker, 2003).

In turn, our findings indicate that depending on their institutional focus, universities and colleges may need to give appropriate weight to SET ratings when evaluating their professors. Universities and colleges focused on student learning may need to give minimal or no weight to SET ratings. In contrast, universities and colleges focused on students' perceptions or satisfaction rather than learning may want to evaluate their faculty's teaching using primarily or exclusively SET ratings, emphasize to their faculty members the need to obtain as high SET ratings as possible (i.e., preferably the perfect ratings), and systematically terminate those faculty members who do not meet the standards. For example, they may need to terminate all faculty members who do not exceed the average SET ratings of the department or the university, the standard of satisfactory teaching used in some departments and universities today despite common sense objections that not every faculty member can be above the average.

Acknowledgements

We thank Laura Grant and Kelsey Cnudde for assistance with preparation of this article.

References

Abrami, P. C., & d'Apollonia, S. (1999). Current concerns are past concerns. *American Psychologist*, *54*(7), 519–520. http://dx.doi.org/10.1037/0003-066X.54.7.519.

Abrami, P. C., Cohen, P. A., & d'Apollonia, S. (1988). Implementation problems in meta-analysis. *Review of Educational Research*, *58*(2), 151–179. http://dx.doi.org/10.3102/00346543058002151.

Abrami, P. C., D'Apollonia, S., & Cohen, P. A. (1990). Validity of student ratings of instruction: What we know and what we do not. *Journal of Educational Psychology*, *82*(2), 219–231. http://dx.doi.org/10.1037/0022-0663.82.2.219.

Beleche, T., Fairris, D., & Marks, M. (2012). Do course evaluations truly reflect student learning? Evidence from an objectively graded post-test. *Economics of Education Review*, *31*(5), 709–719. http://dx.doi.org/10.1016/j.econedurev.2012.05.001.

Benbassat, J., & Bachar, E. (1981). Validity of students' ratings of clinical instructors. *Medical Education*, *15*(6), 373–376.

Bendig, A. W. (1953a). Student achievement in introductory psychology and student ratings of the competence and empathy of their instructors. *The Journal of Psychology: Interdisciplinary and Applied*, *36*, 427–433. http://dx.doi.org/10.1080/00223980.1953.9712910.

Bendig, A. W. (1953b). The relation of level of course achievement to students' instructor and course ratings in introductory psychology. *Educational and Psychological Measurement*, *13*, 437–448. http://dx.doi.org/10.1177/001316445301300307.

Benton, S. L., & Cashin, W. E. (2012). *Student ratings of teaching: A summary of research and literature*. IDEA Center Kansas State University1–19.

Benton, S. E., & Scott, O. (1976). A Comparison of the Criterion Validity of Two Types of Student Response Inventories for Appraising Instruction. *A paper presented at the Annual Meeting of the National Council on Measurement in Education*.

Bolton, B., Bonge, D., & Marr, J. (1979). Ratings of instruction, examination performance, and subsequent enrollment in psychology courses. *Teaching of Psychology*, *6*(2), 82–85. http://dx.doi.org/10.1207/s15328023top0602_6.

Borg, W. R., & Hamilton, E. R. (1956). Comparison between a performance test and criteria of instructor effectiveness. *Psychological Reports*, *2*, 111–116. http://dx.doi.org/10.2466/PR0.2.3.

Braskamp, L. A., Caulley, D., & Costin, F. (1979). Student ratings and instructor self-ratings and their relationship to student achievement. *American Educational Research Journal*, *16*(3), 295–306. http://dx.doi.org/10.2307/1162782.

Bryson, R. (1974). Teacher evaluations and student learning: A reexamination. *Journal of Educational Research*, *68*(1), 12–14.

Capozza, D. R. (1973). Student evaluations, grades, and learning in economics. *Western Economic Journal*, *11*(1), 127.

Centra, J. A. (1977). Student ratings of instruction and their relationship to student learning. *American Educational Research Journal*, *14*(1), 17–24. http://dx.doi.org/10.2307/1162516.

Chase, C.I., Keene, J.M., Jr., Indiana Univ., B.B. of E. S. and T. (1979). Validity of Student Ratings of Faculty. Indiana Studies in Higher Education, Number Forty.

Clayson, D. E. (2009). Student evaluations of teaching: Are they related to what students learn? A meta-analysis and review of the literature. *Journal of Marketing Education*, *31*(1), 16–30. http://dx.doi.org/10.1177/0273475308324086.

Cohen, S. H., & Berger, W. G. (1970). Dimensions of students' ratings of college instructors underlying subsequent achievement on course examinations. *Proceedings of the annual convention of the american psychological association*, *5*, 605–606 [Pt. 2].

Cohen, P. A. ([449_TD$DIFF]1980). A meta-analysis of the relationship between student ratings of instruction and student achievement. *Unpublished doctoral disertation*. The University of Michigan.

Cohen, P. A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. *Review of Educational Research*, *51*(3), 281–309. http://dx.doi.org/10.2307/1170209.

Cohen, P. A. (1982). Validity of student-ratings in psychology courses – A research synthesis. *Teaching of Psychology*, *9*(2), 78–82. http://dx.doi.org/10.1207/s15328023top0902_3.

Cohen, P. A. (1983). A selective review of the validity of student-Ratings of teaching – comment. *Journal of Higher Education*, *54*(4), 448–458. http://dx.doi.org/10.2307/1981907.

Costin, F. (1978). Do Student Ratings of College Teachers Predict Student Achievement? *Teaching of Psychology*, *5*, 86–88.

Crooks, T., & Smock, H. (1974). *Student ratings of instructors related to student achievement*. University of Illinois: Urbana, Ill : Office of Instructional Resources.

Davis, B. G. (2009). *Tools for teaching*, 2nd ed. San Francisco, CA: Jossey-Bass.

Dowell, D. A., & Neal, J. A. (1982). A selective review of the validity of student ratings of teachings. *The Journal of Higher Education*, *53*(1), 51–62. http://dx.doi.org/10.2307/1981538.

Doyle, K. O., & Crichton, L. I. (1978). Student, peer, and self evaluations of college instructors. *Journal of Educational Psychology*, *70*(5), 815–826. http://dx.doi.org/10.1037/0022-0663.70.5.815.

Doyle, K. O., & Whitely, S. E. (1974). Student ratings as criteria for effective teaching. *American Educational Research Journal*, *11*(3), 259–274. http://dx.doi.org/10.2307/1162199.

Drysdale M.J. (2010, January 1). Psychometric Properties of Postsecondary Students' Course Evaluations. ProQuest LLC.

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*(2), 455–463. http://dx.doi.org/10.1111/j.0006-341x.2000.00455.

Elliott, D.N., 1950. Characteristics and relationships of various criteria of college and university teaching. 70, 5–61.

Ellis, N. R., & Rickard, H. C. (1977). Evaluating the teaching of introductory psychology. *Teaching of Psychology*, *4*(3), 128.

Endo, G. T., & Della-Piana, G. (1976). A validation study of course evaluation ratings. *Improving College and University Teaching*, *24*, 84–86.

Feldman, K. A. (1989). The association between student ratings of specific instructional dimensions and student achievement: Refining and extending the synthesis of data from multisection validity studies. *Research in Higher Education*, *30*(6), 583–645.

Fenderson, B. A., Damjanov, I., Robeson, M. R., & Rubin, E. (1997). Relationship of students' perceptions of faculty to scholastic achievement: Are popular instructors better educators? *Human Pathology*, *28*(5), 522–525. http://dx.doi.org/10.1016/S0046-8177(97)90072-1.

Frey, P. W., Leonard, D. W., & Beatty, W. W. (1975). Student ratings of instruction: Validation research. *American Educational Research Journal*, *12*(4), 435–447.

Frey, P. W. (1973). Student ratings of teaching: validity of several rating factors. *Science*, *182*, 83–85.

Frey, P. W. (1976). Validity of student instructional ratings: Does timing matter? *Journal of Higher Education*, *47*, 327–336.

Galbraith, C. S., & Merrill, G. B. (2012). Predicting student achievement in university-Level business and economics classes: Peer observation of classroom instruction and student ratings of teaching effectiveness. *College Teaching*, *60*(2), 48–55. http://dx.doi.org/10.1080/87567555.2011.627896.

Galbraith, C. S., Merrill, G. B., & Kline, D. M. (2012). Are Student Evaluations of Teaching Effectiveness Valid for Measuring Student Learning Outcomes in Business Related Classes? A Neural Network and Bayesian Analyses. *Research in Higher Education*, *53*, 353–374. http://dx.doi.org/10.1007/s11162-011-9229-0.

Gessner, P. K. (1973). Evaluation of instruction. *Science*, *180*(4086), 566–570. http://dx.doi.org/10.2307/1736154.

Greenwood, G. E., Hazelton, A., Smith, A. B., & Ware, W. B. (1976). A study of the validity of four types of student ratings of college teaching assessed on a criterion of student achievement gains. *Research in Higher Education*, *5*(2), 171–178. http://dx.doi.org/10.1007/BF00992010.

Grush, J. E., & Costin, F. (1975). The student as consumer of the teaching process. *American Educational Research Journal*, *12*(1), 55–66. http://dx.doi.org/10.2307/1162580.

Hoffman, R. G. (1978). Variables affecting university student ratings of instructor behavior. *American Educational Research Journal*, *15*(2), 287–299. http://dx.doi.org/10.2307/1162467.

Johnson, V. E. (2003). *Grade inflation: A crisis in college education*. New York: Springer.

Koon, J., & Murray, H. (1995). Using multiple outcomes to validate student-ratings of overall teacher-effectiveness. *Journal of Higher Education*, *66*(1), 61–81. http://dx.doi.org/10.2307/2943951.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*(6), 1121–1134. http://dx.doi.org/10.1037/0022-3514.77.6.1121.

Marsh, H. W., & Overall, J. U. (1980). Validity of students' evaluations of teaching effectiveness: Cognitive and affective criteria. *Journal of Educational Psychology*, *72*(4), 468–475. http://dx.doi.org/10.1037/0022-0663.72.4.468.

Marsh, H. W., Fleiner, H., & Thomas, C. S. (1975). Validity and usefulness of student evaluations of instructional quality. *Journal of Educational Psychology*, *67*(6), 833–839. http://dx.doi.org/10.1037/0022-0663.67.6.833.

Marsh, H. W. (2007). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness. In R. P. Perry, & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 319–383).Netherlands: Springer Retrieved from: http://link.springer.com.ezproxy.library.ubc.ca/chapter/10.1007/1-4020-5742-3_9.

McCallum, L. W. (1984). A meta-analysis of course evaluation data and its use in the tenure decision. *Research in Higher Education*, *21*(2), 150–158. http://dx.doi.org/10.1007/BF00975102.

McKeachie, W. J., et al. (1971). Student ratings of teacher effectiveness: Validity studies. *American Educational Research Journal*, *8*, 435–445.

McKeachie, W. J., Lin, Y., & Mendelson, C. N. (1978). A small study assessing teacher effectiveness: Does learning last? *Contemporary Educational Psychology*, *3*(4), 352–357. http://dx.doi.org/10.1016/0361-476X(78)90037-1.

Mintzes, J. J. (1977). Field test and validation of a teaching evaluation instrument: The student opinion survey of teaching. *A report submitted to the senate committee for teaching and learning, faculty senate*. Windsor, Ontario: University of Windsor.

Moreno, S. G., Sutton, A. J., Ades, A., Stanley, T. D., Abrams, K. R., Peters, J. L., et al. (2009). Assessment of regression-based methods to adjust for publication bias through a comprehensive simulation study. *BMC Medical Research Methodology*, *9*, 2. http://dx.doi.org/10.1186/1471-2288-9-2.

Morgan, W. D., & Vasché, J. D. (1978). An educational production function approach to teaching effectiveness and evaluation. *The Journal of Economic Education*, *9*(2), 123–126.

Morsh, J. E., Burgess, G. G., & Smith, P. N. (1956). Student achievement as a measure of instructor effectiveness. *Journal of Educational Psychology*, *47*(2), 79–88. http://dx.doi.org/10.1037/h0043123.

Murdock R.P., Utah Univ., S.L. C. (1969). The effect of student ratings of their instructor on the student's achievement and rating. Final report. Utah University, Salt Lake City.

Murray, H. G. (1983). Low-inference classroom teaching behaviors and student ratings of college teaching effectiveness. *Journal of Educational Psychology*, *75*(1), 138–149. http://dx.doi.org/10.1037/0022-0663.75.1.138.

Murray, H. G. (2005). Student evaluation of teaching: Has it made a difference? *Presented at the annual meeting of the society for teaching and learning in higher education* Retrieved from: https://www.stlhe.ca/wp-content/uploads/2011/07/Student-Evaluation-of-Teaching1.pdf.

Orpen, C. (1980). Student evaluation of lecturers as an indicator of instructional quality: A validity study. *Journal of Educational Research*, *74*(1), 5–7.

Ory, J. C., & Ryan, K. (2001). How do student ratings measure up to a new validity framework? *New Directions for Institutional Research*, *2001*(109), 27.

Palmer, J. (1978). Leniency, learning, and evaluations. *Journal of Educational Psychology*, *70*(5), 855–863.

Prosser, M., & Trigwell, K. (1991). Student evaluations of teaching and courses: Student learning approaches and outcomes as criteria of validity. *Contemporary Educational Psychology*, *16*(3), 293–301. http://dx.doi.org/10.1016/0361-476X(91)90029-K.

Rücker, G., Schwarzer, G., Carpenter, J. R., Binder, H., & Schumacher, M. (2011). Treatment-effect estimates adjusted for small-study effects via a limit meta-analysis. *Biostatistics (Oxford England)*, *12*(1), 122–142. http://dx.doi.org/10.1093/biostatistics/kxq046.

Rankin, E. F., Greenmun, R., & Tracy, R. J. (1965). Factors related to student evaluations of a college reading course. *Journal of Reading*, *9*(1), 10–15.

Remmers, H. H., Martin, F. D., & Elliot, D. N. (1949). Are students' ratings of instructors related to their grades? *Purdue University Studies in Higher Education*, *66*, 17–26.

Reynolds, D. V., & Hansvick, C. (1978). Graduate instructors who grade higher receive lower evaluations by students. *In annual meeting of the american psychological association*.

Rodin, M., & Rodin, B. (1972). Student evaluations of teachers. *Science*, *177*, 1164–1166.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2006). *Publication bias in meta-analysis: Prevention assessment and adjustments*. John Wiley & Sons.

Rubenstein, J., & Mitchell, H. (1970). Feeling free, student involvement, and appreciation. *In proceedings of the 78th annual convention of the american psychological association*, *5*, 623–624.

Sheets, D. F., Topping, E. E., & Hoftyzer, J. (1995). The relationship of student evaluations of faculty to student performance on a common final examination in the principles of economics courses. *Journal of Economics (MVEA) 21*(2), 55–64. http://www.cba.uni.edu/economics/joe.htm.

Shmanske, S. (1988). On the measurement of teacher effectiveness. *The Journal of Economic Education*, *19*(4), 307–314. http://dx.doi.org/10.1080/00220485.1988.10845278.

Solomon, D., Rosenberg, L., & Bezdek, W. E. (1964). Teacher behavior and student learning. *Journal of Educational Psychology*, *55*(1), 23–30. http://dx.doi.org/10.1037/h0040516.

Soper, J. C. (1973). Soft research on a hard subject: Student evaluations reconsidered. *Journal of Economic Education*, *5*, 22–26.

Sorge, D. H., & Kline, C. E. (1973). Verbal behavior of college instructors and attendant effect upon student attitudes and achievement. *College Student Journal*, *7*(4), 24–29.

Spencer, R. E., & Dick, W. (1965). *Course evaluations questionnaire: Manual of interpretation (Research report No. 200)*. Urbana, I11: Office of Instructional Resources, University of Illinois.

Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the validity of student evaluation of teaching the state of the art. *Review of Educational Research*, *83*(4), 598–642. http://dx.doi.org/10.3102/0034654313496870.

Stanley, T. D., & Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods*, *5*(1), 60–78. http://dx.doi.org/10.1002/jrsm.1095.

Sullivan, A. M., & Skanes, G. R. (1974). Validity of student evaluation of teaching and the characteristics of successful instructors. *Journal of Educational Psychology*, *66*(4), 584–590. http://dx.doi.org/10.1037/h0036929.

Turner, R.L., Thompson, R.P., 1973. Relationships Between College Student Ratings of Instructors and Residual Learning. A paper presented at the annual meetings of the Amerian Educational Research Association, Chicago, Illinois..

Uttl, B., Eche, A., Fast, O., Mathison, B., Valladares Montemayor, H., & Raab, V. (2012). *Student evaluation of instruction/teaching (SEI/SET) review*. Calgary, AB, Canada: Mount Royal Faculty Association Retrieved from:http://mrfa.net/files/MRFA_SEI_Review_v6. pdf.

Uttl, B., White, C. A., & Morin, A. (2013). The numbers tell it all: Students don't like numbers. *Plos One*, *8*(12) . http://dx.doi.org/10.1371/journal.pone.0083443 e83443-e83443.

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science*, *4*(3), 274–290. http://dx.doi.org/10.1111/j.1745-6924.2009.01125.x.

Wachtel, H. K. (1998). Student evaluation of college teaching effectiveness: A brief review. *Assessment & Evaluation in Higher Education*, *23*(2), 191.

Weinberg, B. A., Hashimoto, M., & Fleisher, B. M. (2009). Evaluating teaching in higher education. *Journal of Economic Education*, *40*(3), 227–261.

Wherry, R. J. (1952). *Control of bias in ratings (PRS reports, 914, 915, 919, 920, 921)*. Washington, D.C: Deparment of Army, The Adjutant General's Office.

Whitely, S. E., & Doyle, K. O. Jr. (1979). Validity and generalizability of student ratings from between-classes and within-class data. *Journal of Educational Psychology*, *71*, 117–124.

Wiviott, S. P., & Pollard, D. S. (1974). Background, section, and student evaluation variables as predictors of achievement in a college course. *The Journal of Educational Research*, *68*(1), 36–42. http://dx.doi.org/10.1080/00220671.1974.10884698.

Yunker, P. J., & Yunker, J. A. (2003). Are student evaluations of teaching valid? Evidence from an analytical business core course. *Journal of Education for Business*, *78*(6), 313–317.