Abstract Title Page

Title: Developing Empirical Benchmarks of Teacher Knowledge Effect Sizes in Studies of Professional Development Effectiveness

Authors and Affiliations:

Geoffrey Phelps
Educational Testing Service
Email: gphelps@ets.org

Nathan Jones (Corresponding Author)
Educational Testing Service
Email: ndjones@ets.org

Ben Kelcey
Wayne State University
Email: ben.kelcey@gmail.com

Shuangshuang Liu
Educational Testing Service
Email: sliu002@ets.org

Zahid Kisa
University of Pittsburgh
Email: zak9@pitt.edu

Background / Context:

Growing interest in teaching quality and accountability has focused attention on the need for rigorous studies and evaluations of professional development programs (Darling-Hammond & Sykes, 1999; The Holmes Group, 1986; National Commission on Teaching and America's Future, 1997). This is particularly true given the consistent evidence that teachers differ substantially in their effectiveness (e.g., Aaronson, Barrow, & Sander, 2007; Rivkin, Hanushek, & Kain, 2005; Rockoff, 2004). Teacher development is increasingly viewed as one of the primary levers for improving teaching quality and ultimately student achievement (Correnti, 2007; Desimone, 2009), which has led major funding agencies to devote substantial resources to measuring and improving teacher quality and effectiveness.

However, the study of PD has been hampered by a lack of suitable instruments. Student outcomes are arguably too distal because many factors intervene between effective PD and what students learn (Yoon, Duncan, Lee, Scarloss, and Shapley, 2007). Conventional assessments of teacher content knowledge typically consist of straight subject matter tests and do not focus on the specialized types of content knowledge emphasized in PD and used in teaching (Ball, Thames & Phelps, 2010). Proxies such as teachers self-reports of their knowledge or learning do not assess what teachers actually know or learn (Garet, Porter, Desimone, Birman, & Yoon, 2001; Desimone, Porter, Garet, Yoon, & Birman, 2002).

We suggest that a more direct and proximal outcome of PD is teacher knowledge, which also serves as an important mediator in explaining effects of PD on student outcomes. There is empirical evidence supporting teachers' knowledge as a critical ingredient and central outcome of effective PD. Recent research has linked PD with changes in teachers' knowledge and quality (Correnti, 2007; Garet et al., 2001). More recent literature has also established links among teacher knowledge and student learning in multiple subjects (Carlisle, et al., 2011; Kelcey, 2011; Hill, Rowan, Ball, 2005). Further, federal policy has acknowledged the importance of teachers' knowledge and its role in teachers' PD (Yoon et al., 2007), and IES has repeatedly identified teachers' knowledge as a valued outcome (e.g., IES Education Research Grants, 2012, p. 19).

Until recently, there had been few instruments suitable for measures the types of teacher knowledge supported by the literature. However, new models of assessment are emerging that provide a direct measure of the range of content knowledge needed to address the content problems that arise in teaching. One prominent example is the set of Learning Mathematics for Teaching (LMT) measures developed by Ball, Hill, and colleagues (Hill, Schilling & Ball, 2004). There is strong evidence that the LMT assessments measure knowledge that is different from conventional tests of mathematics, specialized to teaching, sensitive to PD treatments, and associated with instructional quality and student outcomes (Hill et al., 2008; Rockoff, Jacob, Kane, Staiger, 2011; Hill, Rowan & Ball, 2004; Hill, Schilling & Ball, 2004).

In this session, we present data from Teacher Knowledge Assessment System (TKAS), which is designed to administer LMT measures. TKAS is being widely adopted in the evaluation of PD programs with over 500 separate program administrations and 16,000 teachers representing every major region in the country. TKAS provides a first of its kind data base that can be used to assess the suitability of teacher knowledge assessments as tools for studying teacher development across a wide range contexts, teachers and program designs.

Purpose / Objective / Research Question / Focus of Study:

The purpose of the current study is to leverage the TKAS dataset to develop a set of empirical benchmarks of effect sizes for designing rigorous studies of teacher professional development programs. While several studies have been conducted to understand how effect sizes vary in PDs of different design features, the existing empirical research base has offered only limited guidance because evidence has been mostly derived from small-scale evaluations of single-site programs with questionable outcome measures (Borko, 2004; Wayne, Yoon, Zhu, Cronen, & Garet, 2008). This paper overcomes these limitations by drawing on the TKAS data for programs serving in-service teachers.

We will use the pre- to post-test gains in knowledge made by teachers over the course of their respective professional development programs as an empirical benchmark. We recognize that these gains include many forms of selection bias and that more precise treatment and control contrasts would be more optimal in indexing professional development impacts. However, the extant literature has offered little in the way of any type of benchmarks of professional development effects (Wayne et al. 2008). As a second step, we present empirical estimates of effect sizes for programs classified according to major design features of PD sessions. These include program duration, program density (contact hours/duration) and characteristics of program participants including starting levels of knowledge and professional attainment.

Setting and Population / Participants / Subjects:

The TKAS database includes 5 different elementary and middle school outcomes: (1) Elementary number and operations; (2) Elementary patterns, functions and algebra; (3) grade 4-8 geometry; (4) Middle school number and operations; (5) Middle school algebra; and (6) early elementary reading. Data come from 41 states and the District of Columbia. While not nationally representative (or representative of states), these data comprise one of the largest samples of teacher PD programs to date.

The TKAS data system is employed by a variety of users (e.g., teacher educators, district personnel) for a variety of purposes. To ensure that our analytic sample only included teachers enrolled in professional development programs, we limited our sample in a few important ways. First, we dropped preservice teachers and preservice programs from the sample. We also excluded PD programs of fewer than 10 teachers, given concerns about the estimates of program effects drawn from such small samples. Finally, we only included teachers who had both pre- and post- data available. The final sample included 5,318 teachers in 259 programs. We present descriptive statistics in Table 1.

Research Design

Our research design consists of two research goals. In the first, we investigate the average effect sizes of over 259 PD programs in math, paying specific attention to variation across programs and variation across outcome measures. To calculate our effect sizes, we follow the convention proposed by Lipsey and Wilson (2000) for one sample pre-post designs; i.e., the pre-group mean is subtracted from the post mean and divided by the SD at pre. These effects sizes are then combined in a random effects meta analysis. We choose to consider professional development programs as random effects because our goal is not to draw inferences to any specific or single professional development program. Rather, we consider our sample of over 250 professional development programs as a sample of the population of professional development programs.

Second, we will leverage the extensive professional development questionnaire describing the enacted professional development interventions and teachers using data collected from surveys of teachers and program providers (i.e., the individuals who led the PD programs in individual sites). Based on our initial review of the literature, we will use the survey data to focus on five professional development characteristics that set programs apart: (1) the pedagogical focus of the professional development (e.g., improving teachers' content knowledge, improving instructional practices, learning how students' think mathematically, how to construct student assessments, etc.), (2) the density (frequency and duration) of the professional development program, (3) the substantive focus of the professional development (e.g., number sense, fractions, geometry, algebra, etc.), (4) the engagement with artifacts directly related to the mathematics instruction (e.g., students' work, videos of instruction), and (5) the teacher reported alignment of the professional development program with their teaching context (e.g., with local standards, curricula and instructional goals).

Please note that in this abstract, we only present findings related to the first research goal, as we are in the process of preparing the survey data for analysis. The final paper presented at SREE will include all of the above analyses.

Significance / Novelty of study:

Our empirical benchmarks of PD effect sizes serve two purposes. For one, such benchmarks provide a context for interpreting the magnitude of effect sizes drawn from future empirical studies. In this sense, our results provide a more relevant point of reference than two other commonly-employed benchmarks—Cohen’s guidelines for small, medium, and large effect sizes, and empirical benchmarks that have been derived from studies using student outcomes. If studies include teacher knowledge as an outcome, the magnitude of their effects should be interpreted relative to other studies of teacher knowledge.

Second, our results will provide guidance in the design of future studies of the effectiveness of PD interventions. As Desimone (2009) and others have argued, teacher knowledge is an important outcome in PD interventions, and it may provide a more proximal measure of a program’s effectiveness than student achievement.

Statistical, Measurement, or Econometric Model:

As described above, to provide an overall description of the distribution of effects from our sample of professional development programs, we calculate effects for the pre-post gains of each program. We then use a random effects model to combine weighted means (using inverse variance weights) of the program effects. This procedure is repeated across each of the five study outcomes.

To assess pre- to post- test gains and the extent to which these gains vary by teacher, school and professional development factors, we will use a cross-classified random effects model. Each teacher in this dataset is cross-classified by school membership and professional development program. As a result, we first use random effects to address the clustering among teachers within schools and, second, use random effects for the clustering of teachers within professional development programs. We consider estimates over the entire sample using

$$\Delta Y_{ijkl}^{(A)} = \gamma_{0000} + u_{jk} + r_k + t_l + \varepsilon_{ijkl} \tag{1}$$

where $\Delta Y_{ijkl}^{(A)}$ is the knowledge gain for MKT outcome A for teacher i in school j in district k who participated in professional development program l, γ_{0000} is the overall average gain, ε_{ijkl} are the teacher specific residuals (with distribution $N(0,\sigma_{\varepsilon}^2)$), u_{jk} are school specific random effects (with distribution $N(0,\sigma_{\pi}^2)$), r_k are the district specific random effects (with distribution $N(0,\sigma_{\beta}^2)$), and t_l are the PD specific random effects (with distribution $N(0,\sigma_t^2)$).

Findings / Results:

The pre-post effect size estimates are summarized in Table 2, and we also present the number of teachers (N), programs (J), and the calculated unconditional ICCs for the five math outcomes. The effect sizes vary from 0.12 for MSPFA to 0.25 for ELGEO, with an overall mean of 0.19. Across the board, our estimates are smaller than benchmarks that have been developed for measures of student achievement. Hill, Bloom, Black, and Lipsey (2007), for example, conducted a meta-analysis of 61 similar interventions that used student achievement outcomes. They found that for elementary school studies, middle school studies, and high school studies, the average effect sizes were 0.33, 0.51, and 0.27 respectively. Our results suggest that these student achievement benchmarks may be inappropriate for the design of studies that use teacher knowledge as an outcome. Also, it is worth noting that there is some meaningful variation attributable to programs, with ICCs ranging from .01 to .16.

As an example, in Figure 1, we present the distribution of effect sizes (and their confidence intervals) for one outcome, ELGO, sorted from smallest to largest effect size estimate. These results again highlight just how much variance there is in effect sizes across programs.

While we have conducted preliminary analyses to demonstrate variation in program effect sizes, we are currently evaluating the most defensible ways to categorize programs by design features such as duration, density, content focus, and initial knowledge of the teachers served in the program. Additionally, the final results presented at SREE will also reflect our ongoing efforts to handle missing data from pre- to post-test, as well as the shared variance in outcome test measure for programs that assessed teachers using more than one CKT assessment.

Usefulness / Applicability of Method:

Our sample is the largest to date of these kinds of teacher knowledge measures, affording the opportunity to estimate design parameters of effect sizes for planning group randomized trials in teacher PD (and for evaluating the relative magnitude of effects found in future empirical studies). At the same time, we recognize the limitations of our estimates. All were drawn from pre-post designs and we have not controlled for how teachers were selected into PD programs. With this in mind, one important topic for discussion with the SREE audience will be to discuss the significance of our findings. Given that we lack any empirical benchmarks for interpreting teacher knowledge effect estimates, how useful is to reference ones that may be compromised?

Conclusions:

These findings, though preliminary, provide potentially important guidance for the design of group randomized trials using teacher knowledge as an outcome measure. Most notable about our findings is the variation of effects between and across outcomes, suggesting that researchers should consider the specific math outcome that is most relevant for their intervention. For example, programs with a substantive focus on elementary geometry have an average effect size of 0.25, while those focusing on elementary patterns, functions, and algebra have an average effect of .17. These results provide context relevant guidelines (e.g., specific to substantive focus) that researchers may use individually or in combination to design studies and/or assess the practical impact of teacher PD interventions. At the same time, our estimation of effect sizes will also assist researchers in the interpretation of observed effect sizes from professional development studies by providing relevant and substantively meaningful benchmarks. In general, the effect size estimates drawn from studies using teacher outcomes are smaller than those found in studies using student achievement as an outcome; they are also “small” based on the benchmarks put forward by Cohen. These findings suggest that studies using teacher knowledge as an outcome should be considered differently than those using measure of student achievement. Lastly, as we conduct the analyses to be included in the full paper, we will attend to how certain characteristics of PD programs are associated with effect size estimates. These analyses of the context of PD will greatly improve our estimates of empirical benchmarks of teacher knowledge.

Appendices

Appendix A. References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics, 25*(1), 95-136

Ball, D.L., Thames, M., & Phelps, G. (2008). Content Knowledge for Teaching: What Makes It Special? *Journal of Teacher Education,* 59(5), 389-407.

Bloom, H., Hill, C., Black, A., & Lipsey, M. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. MDRC working papers on research methodology.

Borko, H. (2004). PD and teacher learning: Mapping the terrain. *Educational Researcher,* 33(8), 3–15.

Carlisle, J., Kelcey, B., Rowan, B., & Phelps, G. (2011). Teachers' Knowledge About Early Reading: Effects on Students' Gains in Reading Achievement, *Journal for Research on Educational Effectiveness, 4,* pp. 289-321

Darling-Hammond, L., & Sykes, G. (Eds.). (1999). Teaching as the learning profession: Handbook of policy and practice. San Francisco: Jossey-Bass.

Desimone, L., (2009). Improving Impact Studies of Teachers’ PD: Toward Better Conceptualizations and Measures. *Educational Researcher, 38,* 3, pp. 181-199.

Desimone, L., Porter, A.C., Garet, M., Yoon, K. S., & Birman, B. (2002). Effects of professional development on teachers’ instruction: Results from a three-year study. *Educational Evaluation and Policy Analysis, 24*(2), 81-112.

Garet, M. S., Porter, A. C., Desimone, L. M., Birman, B., & Yoon, K. S. (2001). What makes professional development effective? Analysis of a national sample of teachers. *American*

Hill, H.C., Rowan, B., & Ball, D. (2005). Effects of Teachers’ Mathematical Knowledge for Teaching on Student Achievement. *American Educational Research Journal, Vol. 42, No. 2, pp. 371–406.*

Hill, H. C., Schilling, S., & Ball, D. L. (2004). Developing measures of teachers' mathematics knowledge for teaching. *Elementary School Journal, 105*(1), 11-30.

Hill, H. C., Blunk, M. L., Charalambous, Y., Lewis, J. M., Phelps, G., Sleep, L., & Ball, D. L. (2008). Mathematical Knowledge for Teaching and the Mathematical Quality of Instruction: An Exploratory Study. Cognition and Instruction, 26, 430-511.

The Holmes Group. (1986). Tomorrow’s teachers: A report of the Holmes Group. East Lansing, MI.

Kelcey, B. (2011). Assessing the Effects of Teachers’ Reading Knowledge on Students’ Achievement Using Multilevel Propensity Score Stratification, Educational Evaluation & Policy Analysis, 33, pp. 458-482.

National Commission on Teaching and America’s Future. (1997). Doing what matters most: Investing in quality teaching. New York.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (March, 2005). Teachers, schools, and academic achievement. *Econometrica*, *73*(2), 417-458.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review Papers and Proceedings*, 247-252.

Rockoff, J. E., Jacob, B. A., Kane, T. J., & Staiger, D. O. (2011). Can you recognize an effective teacher when you recruit one? . [Article]. Education Finance and Policy, 6(1), 43-74. doi: 10.1162/EDFP_a_00022

Wayne, A. J., Yoon, K. S., Zhu, P., Cronen, S., & Garet, M. S. (2008). Experimenting with teacher PD: Motives and methods. Educational Researcher, 37(8), 469–479.

Yoon, K.S., Duncan, T., Lee, S. W.-Y., Scarloss, B., and Shapley, K. (2007). Reviewing the Evidence on How Teacher PD Affects Student Achievement (Issues & Answers Report, No. 033). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southwest.

Appendix B. Tables and Figures

Not included in page count.

Table 1:
Selected descriptive statistics for the teachers in the TKAS dataset

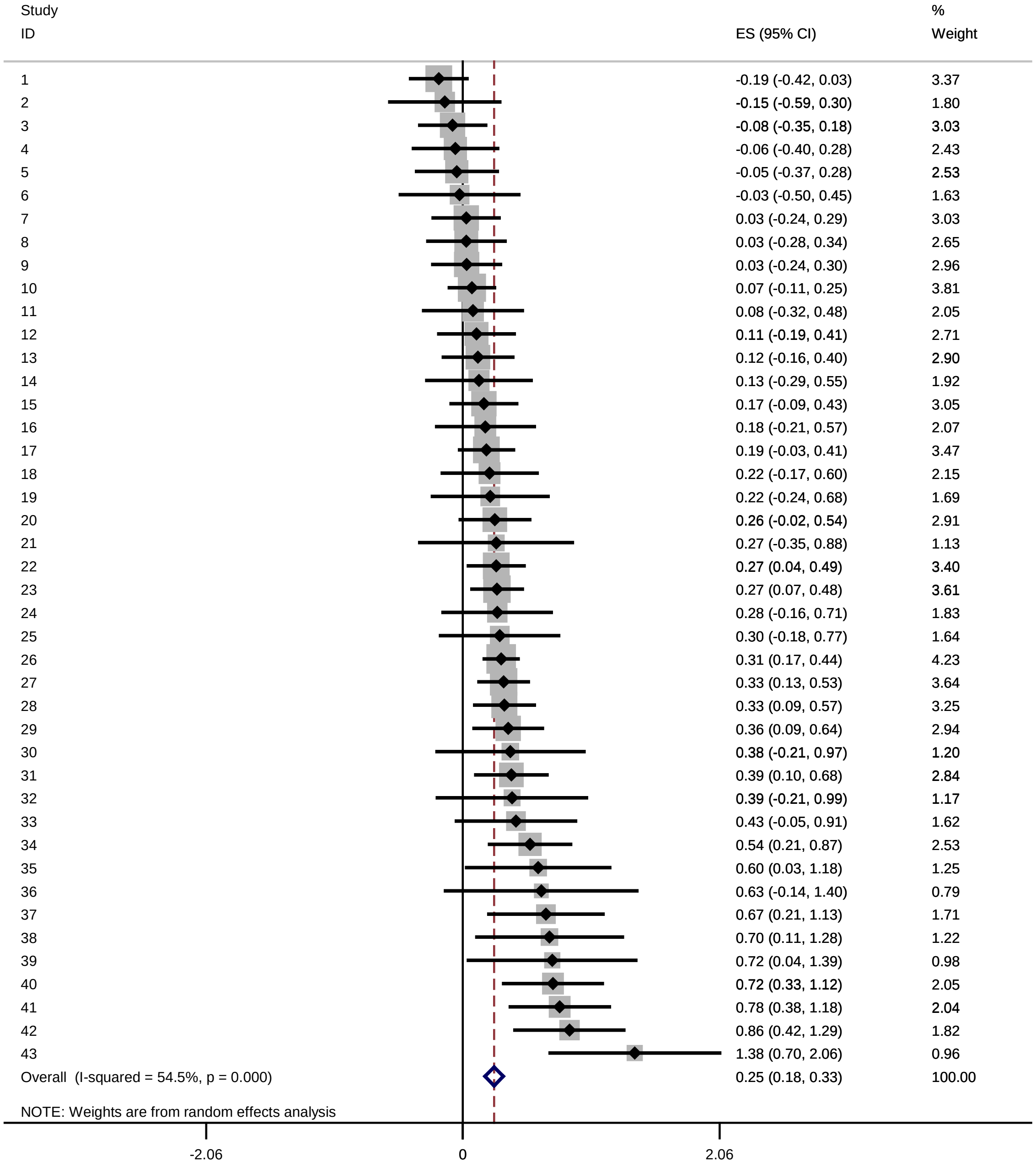
Variable	**Obs**	**Mean**	**Std. Dev.**
# Math Classes	3078	6.93	7.72
# Teaching Math Classes	3047	2.57	5.39
National Board Certified	2963	0.14	0.35
Math Credential K-5	2979	0.64	0.48
Math Credential 6-8	2979	0.68	0.47
Math Credential 9-12	2979	0.34	0.47
Years Teaching Math	2967	10.45	8.40
Teaching Math K-2	2979	0.18	0.38
Teaching Math 3-5	2979	0.31	0.46
Teaching Math 6-8	2979	0.41	0.49
Teaching Math 9-12	2979	0.23	0.42
Female	3091	1.21	0.41
White	3077	0.71	0.43

Note: These descriptive statistics are based on teachers who completed surveys at pre- and post- survey administration and thus do not reflect the full analytic sample.

Table 2
Average Program Effect Size (by Math Outcome)

	N	J	ICC	Mean ES	Mean SE
Elem . Geometry	1170	43	0.10	0.25	0.18
Elem. Number Concepts & Operations	2029	69	0.09	0.21	0.13
Elem. Patterns, Functions, & Algebra	1362	47	0.01	0.17	0.11
Mid. Sch. Number Concepts & Operations	856	33	0.16	0.24	0.14
Mid. Sch. Patterns, Functions, & Algebra	2297	63	0.09	0.12	0.07

Figure 1: *Random Effect Estimates of Program Effect Estimates (ELGEO)*


Study
ID
ES (95% CI)
Weight
1 -0.19 (-0.42, 0.03) 3.37
2 -0.15 (-0.59, 0.30) 1.80
3 -0.08 (-0.35, 0.18) 3.03
4 -0.06 (-0.40, 0.28) 2.43
5 -0.05 (-0.37, 0.28) 2.53
6 -0.03 (-0.50, 0.45) 1.63
7 0.03 (-0.24, 0.29) 3.03
8 0.03 (-0.28, 0.34) 2.65
9 0.03 (-0.24, 0.30) 2.96
10 0.07 (-0.11, 0.25) 3.81
11 0.08 (-0.32, 0.48) 2.05
12 0.11 (-0.19, 0.41) 2.71
13 0.12 (-0.16, 0.40) 2.90
14 0.13 (-0.29, 0.55) 1.92
15 0.17 (-0.09, 0.43) 3.05
16 0.18 (-0.21, 0.57) 2.07
17 0.19 (-0.03, 0.41) 3.47
18 0.22 (-0.17, 0.60) 2.15
19 0.22 (-0.24, 0.68) 1.69
20 0.26 (-0.02, 0.54) 2.91
21 0.27 (-0.35, 0.88) 1.13
22 0.27 (0.04, 0.49) 3.40
23 0.27 (0.07, 0.48) 3.61
24 0.28 (-0.16, 0.71) 1.83
25 0.30 (-0.18, 0.77) 1.64
26 0.31 (0.17, 0.44) 4.23
27 0.33 (0.13, 0.53) 3.64
28 0.33 (0.09, 0.57) 3.25
29 0.36 (0.09, 0.64) 2.94
30 0.38 (-0.21, 0.97) 1.20
31 0.39 (0.10, 0.68) 2.84
32 0.39 (-0.21, 0.99) 1.17
33 0.43 (-0.05, 0.91) 1.62
34 0.54 (0.21, 0.87) 2.53
35 0.60 (0.03, 1.18) 1.25
36 0.63 (-0.14, 1.40) 0.79
37 0.67 (0.21, 1.13) 1.71
38 0.70 (0.11, 1.28) 1.22
39 0.72 (0.04, 1.39) 0.98
40 0.72 (0.33, 1.12) 2.05
41 0.78 (0.38, 1.18) 2.04
42 0.86 (0.42, 1.29) 1.82
43 1.38 (0.70, 2.06) 0.96
Overall (I-squared = 54.5%, p = 0.000) 0.25 (0.18, 0.33) 100.00
NOTE: Weights are from random effects analysis