Published in final edited form as:
Med Care. 2009 March ; 47(3): 356–363. doi:10.1097/MLR.0b013e3181893f6b.

Meta-Analysis: Audit & Feedback Features Impact Effectiveness on Care Quality

Dr. Sylvia J. Hysong, Ph.D.
Houston VA HSR&D Center of Excellence, Health Services Research & Development Service, Department of Veterans Affairs Medical Center, Houston, TX; Department of Medicine, Baylor College of Medicine, 2002 Holcombe Blvd. (152), Houston, Texas 77030, (713) 794-8616 Phone, (713) 487-7359 Fax

Sylvia J. Hysong: hysong@bcm.tmc.edu

Abstract

Background—Audit and feedback (A&F) has long been used to improve quality of care, albeit with variable results. This meta-analytic study tested whether Feedback Intervention Theory, a framework from industrial/organizational psychology, explains the observed variability in health care A&F research.

Data Source: Studies cited by Jamtvedt's 2006 Cochrane systematic review of A&F, followed by database searches using the Cochrane review's search strategy to identify more recent studies.

Inclusion Criteria: Cochrane review criteria, plus: presence of a treatment group receiving only A&F; a control group receiving no intervention; a quantitatively measurable outcome; minimum n of 10 per arm; sufficient statistics for effect size calculations.

Moderators: Presence of discouragement and praise; correct solution, attainment level, velocity, frequency, and normative information; feedback format (verbal, textual, graphic, public, computerized, group vs. individual); goal setting activity.

Procedure: Meta-analytic procedures using the Hedges-Olkin method.

Results—Of 519 studies initially identified, 19 met all inclusion criteria. Studies were most often excluded due to the lack of a feedback-only arm. A&F has a modest, though significant positive effect on quality outcomes (d=.40, 95% CI ±.20); providing specific suggestions for improvement, written, and more frequent feedback strengthened this effect, whereas graphical and verbal feedback attenuated this effect.

Conclusions—A&F effectiveness is improved when feedback is delivered with specific suggestions for improvement, in writing, and frequently. Other feedback characteristics could also potentially improve effectiveness; however, research with stricter experimental controls is needed to identify the specific feedback characteristics that maximize its effectiveness.

Correspondence to: Sylvia J. Hysong, hysong@bcm.tmc.edu.

Earlier versions of this paper were presented at:

- 8th Annual Health Services and Outcomes Research Conference. Houston, TX: November 6.
- 2008 AcademyHealth Annual Research Meeting. Washington, DC: June 8-10.

Audit and feedback (A&F), that is, furnishing providers with "summaries of clinical performance of health care over a specified period of time(1)" has a longstanding tradition as an intervention to change provider behavior, and consequently, quality of health care. As a form of "knowledge of results(2, 3)", it is thought to improve performance by offering providers current performance information and motivation to improve; A&F has been used to improve a wide range of behaviors in clinical practice across many different settings (4), (5), (6), making it a highly flexible intervention. Though in the past A&F may have been laborious, requiring manual abstraction of paper charts, the increase in providers with access to electronic medical records(7) makes A&F a more feasible proposition. Recently, A&F has gained renewed attention due to its essential role in effectiveness of and attitudes toward emerging physician-based performance measurement and pay-for-performance initiatives (8, 9).; A&F has also been suggested as an important component in continuing education, as research has shown physicians have limited ability to accurately assess their continuing education needs(10). Consequently, health care organizations, providers, and patients alike thus stand to gain significantly from a well designed and implemented A&F intervention.

Despite its potential, research reports that A&F is variably effective.(1, 11, 12) In their systematic review of A&F effectiveness, Jamtvedt and colleagues(1) found mixed results and attributed these findings partially to differences in the specific features of the various feedback interventions.(13) Studies examining specific features of A&F are scarce in the health care literature; one possible reason for this is the lack of a theoretical framework within health care to describe the most impactful components of a feedback intervention. As Foy and colleagues point out,(13) we have "an inadequate understanding of the causal mechanisms by which [A&F] or its variants might exert their effects". Without such a framework, we can neither understand what factors may impact the effectiveness of A&F nor refine the interventions.

Kluger and DeNisi's(14) Feedback Intervention Theory (FIT), a well-documented framework from industrial/organizational psychology, could apply to A&F in health care and may provide the necessary lens through which A&F interventions could be better understood and evaluated. Thus, the purpose of the present research is to apply FIT to the problem of A&F effectiveness in health care settings to help explain observed findings in the health care literature.

Conceptual Model: Understanding How Feedback Works via FIT

In their seminal work, Kluger and DeNisi formulated FIT and presented robust meta-analytic support for its tenets, thus dispelling the then popular belief that all feedback interventions were effective. According to FIT, behavior is regulated by comparing feedback to hierarchically organized goals or standards (e.g., providers drawing blood from a patient do so in the same manner until they notice they are not meeting some standard, such as patients complaining of painful blood draws). Attention is limited and usually directed at a moderate level of the hierarchy; only gaps that receive attention have the potential for change. Thus feedback interventions work by providing new information that redirect recipients' attention either toward or away from the task (i.e. the clinical performance issue in question, such as prescribing appropriate medication). Phenomena that redirect attention

toward the details of the task tend to strengthen feedback's effect on task performance; phenomena that shift attention away from the task tend to weaken this effect(14).

Kluger and DeNisi proposed that three factors determine how effectively this attentional shift occurs: (1) characteristics of the feedback itself (these can be content or format related), (2) the nature of the task performed, and (3) situational and personality variables. This paper focuses on feedback characteristics exclusively, as they seem to be studied in the least detail in the A&F literature (e.g., Jamtvedt already examined task characteristics such as task complexity and outcome seriousness, but examined feedback characteristics as a single "intensity" variable).

In their meta-analytic test of FIT, Kluger and DeNisi found several feedback characteristics that significantly impacted A&F effectiveness, consistent with their propositions: (a) discouragement (providing discouraging verbiage, e.g., "your performance was substandard"), praise (providing encouraging verbiage, e.g., "you are an excellent provider"), and verbally delivered feedback, which directed attention away from the focal task, decreased effects on performance; (b) velocity (amount of change in performance since last feedback intervention), correct solution information (information that helps the feedback recipient see what must change to improve performance, e.g., suggesting appropriate medications and dosages in inappropriate prescribing reports), and feedback delivered via computer, which directed attention toward the details of the task increased feedback's effecton performance; and (c) feedback intervention effects were greater when accompanied with goal setting efforts. The present research aims to replicate Kluger and DeNisi's findings using health care research studies. Consistent with their work, the following effects are hypothesized:

Hypothesis 1. A&F will positively impact provider performance.

Hypothesis 2. A&F characteristics that shift the locus of attention toward the details of the task will augment the effect of A&F on performance.

Hypothesis 3. A&F characteristics that shift the locus of attention away from the task details will attenuate the effect of A&F on performance.

Method

Identification of Studies

Search strategy—The Cochrane Database of Systematic Reviews has published and updated a systematic review on A&F.(1) Since the present study is a replication aiming to apply a theory from one discipline to the body of literature in another discipline, the studies used by the Jamtvedt systematic review (k=122) served as the primary data source for the present study. Additional literature searches were then conducted on PubMed using the same search strategy as the Jamtvedt systematic review to identify any relevant articles published since the Jamtvedt systematic review was conducted (2005 or later). This search yielded an additional 397 studies.

Inclusion criteria—As the primary source of studies was an existing systematic review, the initial inclusion criteria were those used by the Jamtvedt systematic review: randomized controlled trials of the effectiveness of A&F on objectively measured performance of providers in a health care setting or on health care outcomes. Studies using students as participants, or that measured knowledge or test performance were excluded.

Further, the same inclusion criteria as Kluger and DeNisi were used herein as the purpose of this study is to replicate their findings in the health care literature,. Studies must have (a) a group that received only a feedback intervention; (b) a control group that received no intervention; (c) a measurable outcome of performance; (d) a minimum sample size of 10 per condition; and (e) sufficient information for calculating an effect size. A trained coder and the principal investigator independently screened the studies. Inter-rater agreement was 92%; discrepancies were resolved by discussion and consensus.

Operationalization of Moderators

Kluger and DeNisi's work served as the theoretical and operational framework used to identify and code the moderators of A&F effectiveness herein. Kluger and DeNisi's seven significant feedback characteristics plus five characteristics along which A&F interventions commonly vary were coded for this study. These characteristics fall into three types: characteristics about feedback content, feedback format, and feedback frequency. Appendix A lists the values of the included studies along all moderators.

Feedback content—Studies were coded for the presence or absence of (a) correct solution information (b) attainment level (whether participants received actual performance information \), (c) velocity, and (d) normative information (comparison of subject performance levels with that of others or a reference group). Goal setting type (if any) was also coded: (1) difficult and specific goals, (e.g., “decrease your rates of inappropriate prescribing by 10% in 30 days”) (2)“do your best” goals (e.g., “decrease your rates of inappropriate prescribing as much as you can”), or (3) no goals at all.

Feedback format—Six formats were each dummy coded separately: whether feedback was provided (a) verbally, (b) textually, (c) graphically, (d) delivered via computer (as opposed to a live person), (e) delivered publicly, and (f) whether feedback referred to group vs. individual performance.

Feedback frequency—Frequency was coded as the ratio of the number of feedback episodes to the length of the study period in months.

Procedure and Analyses

Studies were coded for the aforementioned moderators independently by the principal investigator and a trained coder. Inter-rater agreement was 83%; as before, disagreements between coders were resolved via discussion and consensus.

Each study reported multiple outcomes; however, to avoid violating the assumption of independence, each study was represented only once. Given the small number of studies in the data set, and the history of mixed findings in terms of A&F's effectiveness, a “proof of

concept" approach was adopted, and thus the largest effect size from each study (whether positive or negative) was included in the meta-analysis. Nonetheless, sensitivity analyses were conducted using the smallest effect size from each study.

Meta-analytic procedures using Hedges and Olkin's method(15) were used to calculate a mean effect size (the standardized mean difference, or *d*) and the 95% confidence interval for the impact of A&F on task performance, using a random effects model. Random effects models tend to yield larger standard errors than fixed effects models when the number of studies is small(16), often resulting in too conservative of an estimate when comparing subgroups, as in moderator analyses. However, FIT suggests that the task of interest moderates the effectiveness of the feedback intervention. As the current study set encompasses a range of outcome measures, it cannot be assumed that the true effect size is the same for all studies (confirmed via a significant Q statistic ($Q=72.55$, $p<.001$)); thus, random effects estimates were used. Additionally, several diagnostic tests were also performed to confirm the stability of the overall effect size estimate. Leave-one-out analyses were conducted to ensure that no single study unduly influenced the estimate. Cumulative analyses were computed for the study to check for biases in *d* due to publication date and sample size. Small study bias was also tested using Egger's regression test; Rosenthal's failsafe N was computed to test for publication bias.

To test for moderator effects, subgroup analyses were performed using fixed effects estimates. As stated earlier, random effects models greatly overestimate the sampling error when the sample size is small; thus the problem is exacerbated when conducting subgroup analyses. Further, as studies within a subgroup are more homogeneous than the overall study set, the problem of underestimation of sampling error is less than in the overall study set(17). Thus fixed effect estimates are appropriate in this case. Subgroups for a moderator whose confidence intervals did not overlap were considered significantly different from each other. Due to the number of included studies, only single-moderator effects were tested. Meta-regression was used to test for the moderating effect of feedback frequency, as this was a continuous variable.

Results

Studies Included in the Meta-Analysis

Figure 1 summarizes the study selection process. The set of included studies from the Jamtvedt review (k=122) and the additional recent article literature search (k=397) yielded at total of 519 candidate studies for inclusion in the present study. After reviewing the abstracts of the articles published since the Jamtvedt review and applying the Jamtvedt review's exclusion/inclusion criteria, (as well as the Kluger and DeNisi criteria if discernible from the abstract), 393 of the 397 studies were excluded. Figure 1 summarizes the reasons studies were excluded from the study; thus a total of 126 studies were evaluated for inclusion into our study. After applying the Kluger and DeNisi criteria to these 126 studies, 16 of these met all inclusion criteria. As seen in the table, the most common exclusion was the lack of a "feedback only" condition; most of these studies employed multifaceted interventions of which A&F was a part (the complete list of studies, by reason for exclusion, is presented in Appendix B). This is consistent with previous research.(11, 13)

Several strategies were employed to increase the number of included studies. First, the RCT requirement was relaxed to include non-RCT studies that included a concurrent control group. However, none of the potentially eligible studies featured a concurrent control group; thus the RCT criterion was retained. Next, the requirement of 10 subjects per arm was relaxed to seven; two studies were eligible for inclusion; however, sensitivity analyses indicated no change in study results, but an increase in error around the parameter estimates; thus the original criterion was retained as well. Finally, 21 studies reported insufficient statistics in their published article with which to calculate effect sizes. Requests for additional statistics were sent to all 21 authors via either mail or email. Three authors responded, increasing the final number of included studies to 19. This study selection process is summarized in Figure 1.

Omnibus Effect Size Test

Figure 2 presents the effect sizes (Cohen's *d*) and 95% confidence intervals for each of the studies included in the meta-analysis, as well as the overall effect size estimate; Table 1 presents results of the omnibus tests, publication bias and sensitivity analyses. As seen from the figure, 15 of the 19 studies exhibited positive effect sizes, though six of these did not differ significantly from zero. The effect size estimate of .40 (95% CI ±.20), suggests that A&F has a modest, though significant effect on the outcome of interest. Leave-one-out analyses showed no significant change in results, with effect sizes ranging from .33 (95% CI = ±.19) to .44 (95% CI = ±.20), suggesting that no one study unduly influenced the results. Cumulative analyses by year (Q=2.44, p=.11) showed no significant biases in *d* due to studies' publication date. Similar analyses by sample size (n), however, indicate a significant, though small positive effect of sample size on *d*, (Q=7.21, p=.007) suggesting that studies with smaller, less stable sample sizes may be slightly less able to detect an A&F effect; thus, the effect size reported above may represent a conservative estimate. Egger's regression test indicated no evidence of small study bias (t=.93, p=.36). Rosenthal's Fail-safe N estimates 246 studies would be needed to nullify these results, well beyond the 110 studies needed (per the 5k+10 rule) to rule out publication bias.

Sensitivity analyses using minimum effect sizes yielded an estimate of .21 (95% CI = ± .14), which does not significantly differ from the original estimate using the largest effect size for each study. Leave-one-out tests and publication bias tests yielded similar results as analyses of the original data set (see Table 1). However, in the original data set, effect sizes for 11 studies were nonsignificant, including five studies with negative effect sizes. This suggests moderators may be present.

Moderator Analyses

Five moderators, attainment level, praise, discouragement, computer-delivered feedback, and goal setting, were not tested due to a lack of studies for comparison (e.g., no studies reported using discouragement; all studies reported using attainment level, thus no comparisons were possible).

Table 2 summarizes the results of the moderators that were tested. Four moderators significantly impacted the effect of A&F on outcomes: correct solution information and

written feedback delivery augmented feedback effectiveness, whereas verbal and graphic feedback delivery attenuated feedback effectiveness; providing both individual- and group-level feedback may be beneficial, though this cannot be confirmed with these data due to the large standard errors and the small sample size. The presence of normative information did not significantly impact A&F effectiveness, nor did public delivery of feedback. Feedback frequency significantly moderated the effectiveness of A&F, such that more frequent feedback augmented A&F effectiveness (B=.07, p=.025). This last result is contrary to Kluger and DeNisi's original findings; they found that greater frequency of feedback decreased the impact of feedback on performance; however, in their study feedback was operationalized as the number of feedback episodes, not accounting for study period length, which could explain this discrepancy.

One potential concern with aforementioned results is that the moderators could be correlated, thereby making individual moderator tests less interpretable. Correlation analyses suggested individual moderator analysis was appropriate; only three moderator pairs were significantly correlated: velocity and graphical delivery (.65), attainment level and correct solution information (-.5), and graphical and textual delivery (-.65). The first correlation is understandable, as charts and graphs are well-suited for the display of change over time, which is what velocity conveys. The latter two correlations are due to the presence of a disproportionate distribution of studies across the cells; nearly all studies reported attainment level, yet only a few studies provided correct solution information. A similar pattern was observed in graphical vs. textual delivery, hence the strong correlation.

Discussion

Organizational and management research has made significant progress in understanding how feedback works, and much of that knowledge can be applied in health care. For example, in order for feedback to be maximally effective, it needs to keep the recipient focused on the task(14); certain feedback characteristics, such as frequency and individualization tend to augment this effect. Indeed, frequent, individualized, and non-punitive feedback has been shown to be effective in helping primary care providers adhere to clinical practice guidelines(18). This meta-analytic study sought to apply industrial/organizational psychology's FIT to health care's A&F literature to clarify the observed variability in findings. Despite the large number of studies excluded due to the lack of proper experimental controls (e.g., lack of control groups and treatment groups that received only A&F, adequate sample sizes, complete reporting of results), a modest, yet significant, positive effect of A&F effectiveness was found overall. Additionally, providing correct solution information in the feedback, providing feedback in writing rather than verbally or graphically, and more frequent feedback augmented A&F's effect on the outcomes of interest. Providing combined group- and individual-level feedback appeared to impact feedback effectiveness, however, definitive conclusions could not be made due to the large error margins around the estimates for this characteristic.

These findings are largely consistent with Kluger and DeNisi's findings in the managerial literature, where they found significant, positive effects of feedback and significant moderation of this effect by correct solution information (positive) and verbal feedback

delivery (negative). The present finding of graphical feedback delivery as a significant moderator (negative) is counterintuitive and contrary to Kluger and DeNisi's predictions. The set of studies using graphical feedback included two of the three studies that did not use textual feedback; additionally, only one of the studies used correct solution information. Thus it is possible that in this case, graphical delivery might be a marker for a generally poorly conceived or implemented intervention, though this is admittedly speculative.

The findings are also consistent with the Jamtvedt review's findings: both studies found a significant, though modest effect of audit and feedback on clinical performance; further, Jamtvedt and colleagues attributed their results largely to differences in the specific features of the feedback used in the studies. The present study is consistent with these findings and goes one step further by identifying specific, theory-based characteristics that improve feedback effectiveness. Particularly, the use of correct solution information is a unique contribution not previously addressed in the health care literature. A recent clinical practice guideline study suggests that in order to be actionable, feedback must be timely, individualized, and meaningful(18). Other research has noted that feedback effectiveness is significantly augmented when paired with goal setting – this would be consistent with FIT's notion that behavior changes due to feedback-standard comparisons (the goal would provide a clear standard for comparison(19). It is plausible that correct solution information provides added meaning and clear standards; this would be consistent with extant research and would provide a new feature of audit and feedback worth additional study.

Limitations

The most obvious limitation in this research is the large number of studies excluded from analysis and the small resulting sample size. However, some exclusions (e.g., no control group) improve the precision of findings, whereas others (e.g. insufficient statistics) have a detrimental effect. The most common reason for excluding a study was the lack of a “feedback only” group (67/296, or 23%). This reflects a trend towards multifaceted interventions fueled by systematic reviews in the late 90's suggesting that multifaceted interventions were more effective(20, 21), though a more recent literature review suggests they are no more effective(22) than single interventions. Including such studies would have introduced significant error in the overall parameter estimate, and would have confounded any observed moderator effects with other, non-feedback related components of the intervention. Conversely, only a small number of studies were excluded due to insufficient reporting of statistics (18/296, or 6%). Further, no evidence of publication or small study bias was found, and sensitivity analyses showed the effect size to still be significant even if the smallest effect sizes are used. Thus, the findings in this study yield useful information, despite the small number of studies included. Nevertheless, this illustrates the importance of evaluating the effectiveness of theory-based quality improvement interventions with stricter experimental controls, and reporting *both* intervention details and statistical findings in greater detail.

Five moderators were untestable because of the unavailability of studies in a given moderator arm (e.g., no studies were available that used goal setting as part of their audit and feedback intervention). In this case, however, this reflects differences in how work is

done in health care versus other industries, rather than flaws in primary study reporting or gaps in the literature. For example, performance measurement in management settings usually consists of supervisory ratings, and feedback involves a face to face conversation rather than reports of countable performance episodes. Consequently, studies exhibiting feedback characteristics such as praise, goal setting, etc. are much more readily available. There may be other format and content cues not tested by Kluger and DeNisi's work that could be more relevant to reports of countable clinical performance episodes -- an area for future research.

Implications

Health care systems have invested many resources in developing interventions to change provider behavior and improve quality of care. A&F has gained increased acceptance as a strategy to change provider behavior and is now a familiar, widely used intervention. This study provides evidence ofsimple changes to A&F interventions that could be implemented immediately (especially at facilities with electronic medical records), and could improve the quality of care.

This study also makes important theoretical contributions. Feedback is recognized by various organizational theories as an important mechanism for regulating both individual and organizational behavior. This research replicates many of the results originally found by Kluger & DeNisi in a completely different literature; such replication enhances the generalizability of FIT as an explanatory framework for feedback effectiveness.

Finally, the present research introduces a much needed theoretical framework to A&F research in health care, and provides a unifying paradigm for what previous studies have found independently: feedback characteristics, such as format(23), timing, and frequency(18), task characteristics such as complexity(1), and situational variables such as goal setting all affect feedback effectiveness. These results help better understand why feedback has been more effective in certain conditions than others, design more sophisticated, refined feedback intervention initiatives for practice, and possibly significantly improve provider practice without investing precious resources in newer, riskier interventions. Correctly(24, 25) designed and implemented A&F interventions could both help improve quality of care directly, and also complement the effectiveness of other quality improvement strategies such as continuing education and pay-for-performance. Future studies could test other portions of this theoretical framework to select the areas of health care for which feedback interventions would be best suited. Future research could also examine optimal ways of providing correct solution information: research tells us that feedback should be non-punitive; (18, 26-29, 29) what does that mean with respect to correct solution information in primarily numerical feedback reports?

Conclusions

Despite mixed findings in previous reviews, A&F could be a reasonably effective tool for changing provider behavior and thus quality of care, if designed correctly. Feedback reports containing specific suggestions for performance improvement, delivered in writing, delivered frequently can noticeably improve A&F effectiveness. Close attention to other

characteristics to improve feedback's meaningfulness to the feedback recipient could result in additional improvements. FIT provides a viable theoretical framework to guide decisions in designing future A&F interventions. However, research with stricter experimental control and more detailed reporting is sorely needed in this area to more precisely identify the specific characteristics most likely to maximize feedback effectiveness and improve quality of care.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

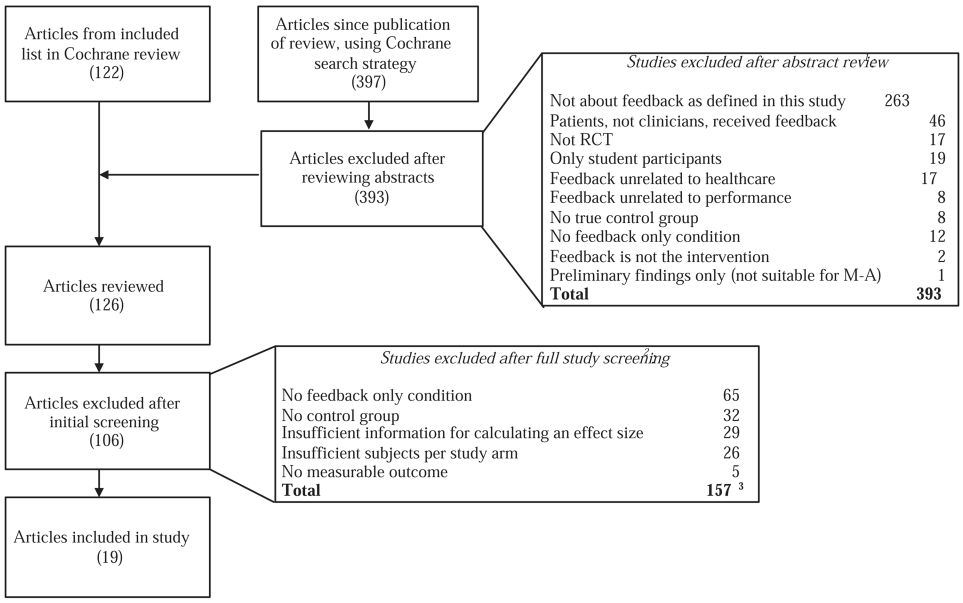
The author would like to gratefully thank Drs. Laura Petersen and Hardeep Singh for their valuable insights and comments in earlier drafts of this manuscript, as well as Dr. Donna White for her methodological insights during the analysis phase, and in earlier drafts of this manuscript. Thanks are also in order to Ms. Joy Malveaux and Mr. Steven Apodaca without whom retrieval and coding of studies would not have been possible.

The research reported here was supported by the Department of Veterans Affairs, Veterans Health Administration, Health Services Research and Development Service (HSR&D) (a Houston Center for Quality of Care & Utilization Studies Locally Initiated Project). The author' salary was supported in part by the Department of Veterans Affairs, and by the National Heart Lung and Blood Institute of the National Institutes of Health (Investigator Research Supplement to Grant no. 1R01HL079173, Laura Petersen, PI). The author declares to have no other competing interests, financial or non-financial. The views expressed in this article are those of the authors and do not necessarily reflect the position or policy of the Department of Veterans Affairs, Baylor College of Medicine, or the National Institutes of Health.

Reference List

1. Jamtvedt G, Young JM, Kristoffersen DT, et al. Audit and feedback: effects on professional practice and health care outcomes. Cochrane Database Syst Rev. 2006; (2):CD000259. [PubMed: 16625533]
2. Karraker RJ. Knowledge of results and incorrect recall of plausible multiple-choice alternatives. Journal of Educational Psychology. 2008; 58(1):11–14.
3. Locke EA. Motivational effects of knowledge of results: Knowledge or goal setting? Journal of Applied Psychology. 1967; 51(4, Pt.1):324–329. [PubMed: 6075575]
4. Bentz CJ, Bayley KB, Bonin KE, et al. Provider feedback to improve 5A's tobacco cessation in primary care: a cluster randomized clinical trial. Nicotine Tob Res. 2007; 9(3):341–349. [PubMed: 17365766]
5. Wahlstrom R, Kounnavong S, Sisounthone B, et al. Effectiveness of feedback for improving case management of malaria, diarrhoea and pneumonia--a randomized controlled trial at provincial hospitals in Lao PDR. Trop Med Int Health. 2003; 8(10):901–909. [PubMed: 14516301]
6. Robling MR, Houston HL, Kinnersley P, et al. General practitioners' use of magnetic resonance imaging: an open randomized trial comparing telephone and written requests and an open randomized controlled trial of different methods of local guideline dissemination. Clin Radiol. 2002; 57(5):402–407. [PubMed: 12014939]
7. Hing ES, Burt CW, Woodwell DA. Electronic medical record use by office-based physicians and their practices: United States, 2006. Adv Data. 2007; (393):1–7.
8. Petersen LA, Woodard LD, Urech T, et al. Does Pay-for-Performance Improve the Quality of Health Care? Annals of Internal Medicine. 2006; 145(4):265–272. [PubMed: 16908917]
9. Meterko M, Young GJ, White B, et al. Provider attitudes toward pay-for-performance programs: development and validation of a measurement instrument. Health Serv Res. 2006; 41(5):1959–1978. [PubMed: 16987311]

10. Davis DA, Mazmanian PE, Fordis M, et al. Accuracy of physician self-assessment compared with observed measures of competence: a systematic review. JAMA. 2006; 296(9):1094–1102. [PubMed: 16954489]
11. Grimshaw JM, Thomas RE, MacLennan G, et al. Effectiveness and efficiency of guideline dissemination and implementation strategies. Health Technology Assessment (Winchester, England). 2004; 8(6):iii–iiv.
12. Thomson O'Brien MA, Oxman AD, Davis DA, et al. Audit and feedback versus alternative strategies: effects on professional practice and health care outcomes. Cochrane Database Syst Rev. 2000; (2):CD000260. [PubMed: 10796521]
13. Foy R, Eccles MP, Jamtvedt G, et al. What do we know about how to do audit and feedback? Pitfalls in applying evidence from a systematic review. BMC health services research. 2005; 5:50. [PubMed: 16011811]
14. Kluger AN, DeNisi A. The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. Psychological Bulletin. 1996; 119(2):254–284.
15. Hedges LV, Olkin I. Clustering estimates of effect magnitude from independent studies. Psychological Bulletin. 1983; 93(3):563–573.
16. Whitehead, A. Meta-analysis of Controlled Clinical Trials: Statistics in Practice. Chichester, England: John Wiley & Sons, Ltd.; 2002.
17. Overton RC. A comparison of fixed-effects and mixed (random-effects) models for meta-analysis tests of moderator variable effects. Psychological Methods. 1998; 3(3):354–379.
18. Hysong SJ, Best RG, Pugh JA. Audit and Feedback and Clinical Practice Guideline Adherence: Making Feedback Actionable. Implementation Science. 2006; 1 serial online. Available at: http://www.implementationscience.com/content/1/1/9.
19. Kluger AN, DeNisi A. Feedback interventions: Toward the understanding of a double-edged sword. Current Directions in Psychological Science. 1998; 7(3):67–72.
20. Grimshaw JM, Shirran L, Thomas R, et al. Changing provider behavior: an overview of systematic reviews of interventions. Medical Care. 2001; 39(8 Suppl 2):II2–45. [PubMed: 11583120]
21. Grol R, Grimshaw J. Evidence-based implementation of evidence-based medicine. Joint Commission Journal on Quality Improvement. 1999; 25(10):503–513. [PubMed: 10522231]
22. Grimshaw J, Eccles M, Tetroe J. Implementing clinical guidelines: current evidence and future implications. J Contin Educ Health Prof. 2004; 24(Suppl 1):S31–S37. [PubMed: 15712775]
23. Mugford M, Banfield P, O'Hanlon M. Effects of feedback of information on clinical practice: a review. BMJ. 1991; 303(6799):398–402. [PubMed: 1912809]
24. Pritchard RD, Jones SD, Roth PL, et al. Effects of group feedback, goal setting, and incentives on organizational productivity. Journal of Applied Psychology. 1988; 73(2):337–358.
25. Dixon E, Hameed M, Sutherland F, et al. Evaluating meta-analyses in the general surgical literature: a critical appraisal. Ann Surg. 2005; 241(3):450–459. [PubMed: 15729067]
26. Hysong SJ, Best RG, Pugh JA, et al. Not of One Mind: Mental Models of Clinical Practice Guidelines in the VA. Health Services Research. 2005; 40(3):823–842.
27. Hysong SJ, Pugh JA, Best RG. Clinical Practice Guideline Implementation Patterns In VHA Outpatient Clinics. Health Serv Res. 2007; 42(1 Pt 1):84–103. [PubMed: 17355583]
28. Kinicki AJ, Prussia GE, Wu B, et al. A covariance structure analysis of employees' response to performance feedback. Journal of Applied Psychology. 2004; 89(6):1057–1069. [PubMed: 15584841]
29. Goldstein, IL.; Ford, JK. Training in Organizations. Belmont, CA: Wadsworth; 2002.



[1] This set of exclusions refers exclusively to the set of articles published since the Cochrane systematic review; since the Cochrane review's criteria were used to search and identify new articles, articles retrieved from the Cochrane review's "included articles" list did not receive abstract review (they went straight to initial screening) and are thus not included here.

[2] This set of exclusions uses all studies retrieved from the Cochrane review's "included articles" list, plus all the articles retained after abstract review.

[3] The total number of studies adds up to more than the total number of studies reviewed (126) because some article failed to meet more than one inclusion criterion. 28 articles failed to meet two exclusion criteria; three articles failed to meet three inclusion criteria, and one article failed to meet all five inclusion criteria.

Figure 1.
Study selection flow diagram.

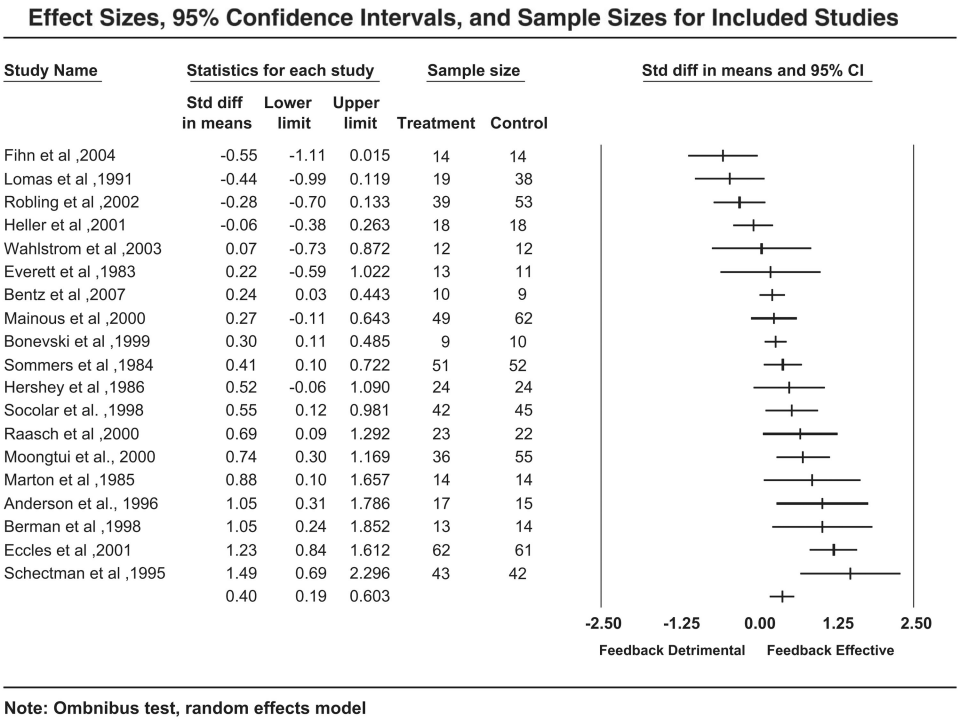


Figure 2.

Effect sizes and 95% confidence intervals of studies included in meta-analysis.

Table 1

Summary of Omnibus Tests, Publication Bias Statistics, and Sensitivity Analyses

Test	Maximum Effect Sizes			Minimum Effect Sizes		
	Omnibus Tests					
		95% Conf. Interval			*95% Conf. Interval*	
	Effect Size (Cohen's d)	***Lower 95% CI Limit***	***Upper 95% CI Limit***	***Effect Size (Cohen's d)***	***Lower 95% CI Limit***	***Upper 95% CI Limit***
Omnibus Test	.40	.19	.60	.21	.07	.35
Leave-one-out						
• Lowest	.33	.14	.51	.18	.04	.31
• Highest	.44	.24	.64	.24	.10	.37
	Tests of Publication Bias					
		Statistic	*p-value*		*Statistic*	*p-value*
• Cumulative by year (Q-statistic)		2.44	.11		.19	.65
• Cumulative by sample size (Q-statistic)		7.21	.01		1.43	.23
• Egger's Regression Test (Student's t)		.93	.36		.22	.82
• Rosenthal's Fail-Safe N		246	N/A		79	N/A

Notes: Effect size reported is Cohen's d, the standardized mean difference between groups.

Table 2

Summary of subgroup analyses for feedback characteristics and meta-regression of feedback frequency on effect size

			95% Conf. Interval	
Moderator	**No. of studies**	**Effect size†**	**LCL**	**UCL**
1. Correct solution information				
Yes*	6	.78[a]	.55	1.00
No*	12	.23[b]	.11	.34
Not Reported*	1	.30 [b]	.11	.48
2. Feedback delivered graphically				
Yes	4	.13[a]	-.05	.31
No*	11	.66[b]	.51	.81
Not Reported	4	.14[a]	-.003	.29
3. Feedback delivered in writing				
Yes*	14	.49[a]	.38	.60
No	3	.10[b]	-.07	.26
Not Reported	2	-.21[b]	-.58	.16
4. Feedback delivered verbally				
Yes	5	.10[a]	-.09	.29
No*	11	.41[b]	.30	.51
Not Reported	3	.25[a,b]	-.06	.57
5. Group vs. individual feedback				
Individual only*	9	.31	.19	.42
Group only*	7	.34	.19	.49
Group and individual*	2	.96	.40	1.52
Not reported	1	.07	-.73	.87
6. Feedback delivered publicly				
Yes*	5	.26	.13	.39
No*	12	.38	.25	.50
Not Reported*	2	.78	.21	1.35
7. Normative information				
Yes*	8	.32	.19	.46
No*	9	.37	.21	.54
Not Reported*	2	.28	.11	.47
Feedback Frequency	*B*††	*SE*	*LCL*	*UCL*
Slope*	.07*	.03	.009	.13
Intercept**	.28**	.05	.18	.38

Notes:

† Effect size reported is Cohen's d, the standardized mean difference between groups.

†† For feedback frequency, reported statistic is the B-weight reflecting the change in Cohen's d per increase in one unit of frequency (i.e. each additional feedback instance results in an estimated increase in effect size of .07).

* Denotes effect is significantly different from zero at the .05 level.

** Denotes effect is significantly different from zero at the .01 level. Within each moderator, subgroups with superscripts of different letters denote subgroups that significantly differ from each other (e.g., studies that used correct solution information have significantly higher effect sizes than the other two subgroups; the other two subgroups, however, do not significantly differ from each other). Subgroups of a moderator without lettered superscripts do not significantly differ from each other.