Supplementary material. Appendix 2

RESULTS: SUPPLEMENTARY INFORMATION

Comparison 1: The effect of verbal face-to-face feedback, compared to no feedback, on performance

Included studies

Participants

Participants included 290 (60%) medical students in four studies,[1-4] 60 (12%) dental students in one study[5] and 138 (28%) doctors (doctors training in surgery in three studies,[6-8] training in obstetrics and gynaecology in one study[9] and training in emergency medicine in one study,[10] and physicians in one study[11]).

Participants were novices to the assessed task in five studies (5/11, 45%);[1, 2, 4, 5, 7] and had prior experience in six studies.[3, 6, 8-11]

Workplace tasks and Settings

All studies evaluated performance of a discrete task; there were no longitudinal evaluations. The task occurred in simulation settings in seven studies (7/11, 64%) and clinical practice in four studies (4/11, 36%). The task was a surgical procedure in seven studies (7/11, 64%). Five studies involved simulated surgical tasks including bench top models for knot tying[4] and forming a bowel anastomosis;[8] using a laparoscopic simulator for suturing and knot tying;[2] and using a virtual reality (VR) simulator for laparoscopic surgery[1] and endovascular surgery.[7] Two studies involved laparoscopic surgery in clinical practice.[6, 9] The remaining four studies evaluated simulated matching of tooth colour in a dental school,[5] simulated cardiopulmonary resuscitation (CPR),[3] chest ultrasound for emergency trauma patients[10] and teaching skills in clinical practice.[11]

Feedback Interventions

The feedback source involved a subject expert in all comparisons except two, including one that compared peer feedback with no feedback,[3] and one that compared expert feedback, peer feedback and no feedback.[7] Feedback occurred while the participant performed the task (during) in one study,[3] both during and directly afterwards in two studies,[1, 2] directly afterwards in four studies,[5, 7, 9, 10] after a delay in three studies[6, 8, 11] and one study compared feedback during, feedback directly afterwards and no feedback.[4] In addition to evaluative performance information (as per inclusion criteria), the feedback included corrective advice in all studies except one[10] and one where it was unclear.[7] Feedback included additional information from a simulator in three studies,[1, 2, 7] a video of the participant's performance in two studies[6, 11] and written performance information in two studies.[5, 11]

Teaching and Practice

In addition, instruction and expert demonstration of the task were provided in six studies (6/11, 55%), including all five studies involving novice participants[1, 2, 4, 5, 7] and one study that involved CPR for medical students, many of whom had previously attended a course.[3] The other five studies involved doctors working in clinical practice; in these studies, no instruction or expert demonstration was included within the research intervention but may or may not have occurred during the course of routine work during that time. One study involved physicians' teaching on ward rounds[11] and the other four studies assessed tasks by doctors training in relevant specialties.[6, 8-10]

The amount of practice varied substantially between different studies, for both simple and complex tasks. For example, comparing two studies that involved simple surgical knot tying: in Xeroulis,[4] participants had 18 practice attempts in one session and in O'Connor,[2] they could practice up to an hour a day, for 24 days. Looking at more complex surgical procedures, such as simulated surgery using a virtual reality (VR) simulator: in Ahlborg,[1] participants had two

practice attempts at the simulated surgery (laparoscopic salpingectomy) and in Boyle,[7] participants had five attempts at the simulated surgery (renal artery angioplasty and stenting) before the performance evaluation.

Intervention period

The intervention period ranged from one day (most common) up to two months.[6] Nine (9/11, 82%) studies involved a single session (involving one episode of feedback in five studies[5, 8-11] and multiple episodes of feedback in four studies[1, 3, 4, 7]). Two studies (2/11, 18%) had a longer intervention period involving multiple feedback sessions: one study[6] included approximately four coaching sessions regarding bariatric surgery across a two month surgical attachment, and another[2] included almost daily one hour practice sessions for laparoscopic suturing, with feedback throughout each one, over four weeks.

The timing of the post-feedback performance assessment, in relation to the intervention, differed. It occurred directly following the intervention in seven studies: at the end of the single session in five studies[1, 3, 4, 7, 9] and at the end of an extended intervention period in two studies.[2, 6] In the other four studies, the post-feedback performance assessment occurred some weeks after the intervention was completed but while relevant exposure to possible teaching and/or practice opportunities continued. Olms[5] included a single feedback session, with the final evaluation two weeks later, in the midst of a routine one month university teaching unit on tooth shade matching. Skeff[11] arranged a single coaching session on ward round teaching in the middle of physicians' four week ward duty, with the final evaluation post-performance evaluation at the end. Soucisse[8] also organised a single coaching session for surgical residents, with the final evaluation occurring three weeks later. Vafaei[10] involved a single workplace-based assessment with feedback for doctors training in emergency medicine on chest ultrasound for emergency trauma patients, followed by a two month period of routine clinical work before the post-feedback assessment.

Research funding

Regarding research funding, one study[3] that focused on cardiopulmonary resuscitation (CPR) quality, was loaned a device (used to measure CPR parameters and provide automated feedback to participants) for the period of the study by Philips but the company was not otherwise involved in the research; five studies received funding from independent institutions,[1, 4, 6, 9, 11] three studies did not receive any funding[5, 7, 10] and two studies did not report information on funding.[2, 8]

Risk of bias

Five trials described an adequate method for randomised sequence generation and allocation concealment, so we rated these studies as 'low risk'.[3, 5, 6, 8, 9] The other six trials simply stated participants were 'randomised' and had no information on allocation concealment, so we rated these studies as 'unclear'. We analysed baseline performance because, although randomisation removes the need to check comparability in baseline task performance for intervention and comparison groups, it may be useful to check this when participant numbers are small and performance improvement is more likely when baseline performance is low.[12] Seven studies reported no statistically significant differences between baseline performances for the comparison groups.[4, 5, 8-11] and four studies did not report baseline task performance.[1-3, 7] The participants and research team members were not blinded in any included studies because the intervention involved feedback between a research team member and a participant, consistent with most education interventions. However, in all included studies, we thought this was not likely to influence the outcome (post-intervention performance assessment) because implementation and adherence to the intervention were not affected. In eight studies the outcome was assessed by either blinded assessors who rated videos of the participants' performance[4, 6-9, 11] or by a machine (simulator or CPR machine),[1, 3] so we rated these as 'low risk' of bias. In three studies, the feedback provider and outcome assessor appeared to be the

same person, so these were rated as 'high risk'.[2, 5, 10] Across all the studies, the follow up rate for each group was at least 85%. Only two studies had a prior published protocol in addition to reporting all outcomes as planned.[6, 8] For all other studies, it could not be ascertained if outcomes had been selectively reported, so these were rated as 'unclear', except one. This one study was rated as 'high risk' for selective outcome reporting because it did not include the expected information on performance post-intervention.[2]

In summarising the risk of bias across domains within each study, two studies had all domains rated 'low risk, so these were rated low risk.[6, 8] Six studies had at least one domain with 'unclear' risk but no 'high risk' ratings, so these were rated as 'unclear' risk of bias.[1, 3, 4, 7, 9, 11] Three studies had at least one domain at high risk of bias, so we judged these studies to be at 'high risk' of bias.[2, 5, 10]

Certainty of evidence

For the comparison of verbal face-to-face feedback compared to no feedback, excluding studies at high risk of bias, we graded the quality of evidence for the outcome of 'objective assessment of a health professional's performance'. The risk of bias was rated as 'unclear' across multiple included studies and the overall body of evidence indicated this was likely to seriously alter the results, so we downgraded the overall evidence by one level. The two aspects that were most influential on our decision were the lack of allocation concealment and prior published protocols to preclude selective reporting of outcomes. Participant and research team member blinding was not possible due to the intervention. However, this had limited impact on the selected outcome 'objective assessment of performance', as no changes occurred in intervention implementation or adherence as a consequence of this lack of blinding.[13] We judged the results to be directly applicable to our review question and therefore the evidence was not downgraded for indirectness. There was some methodological and statistical heterogeneity across studies (the test for heterogeneity was not significant with $P = 0.14$ and I^2

= 34%), which was not explained by subgroup analysis. However, all studies reported a beneficial effect, so the uncertainty seemed to lay in the magnitude of effect rather than the presence of an effect. Therefore, we decided not to downgrade the evidence due to inconsistency.[14] We judged the effect size to be sufficiently precise and therefore did not downgrade the evidence for imprecision of results. This was based on sufficient numbers of participants (392 when studies with high risk of bias were excluded) and a consistent beneficial effect, indicated by the confidence interval for the overall effect estimate not crossing zero and all individual studies showing a beneficial effect with substantial overlap in their confidence intervals. Finally, we judged that there was likely to be a systematic overestimation of the underlying beneficial effect of feedback because we strongly suspected publication bias (see Funnel plot 5b) and therefore we downgraded the evidence by one level.

In summary, combining all five GRADE criteria for assessing the certainty of evidence, we downgraded the overall rating by one, from high to low. We judged that the quality of the evidence was low contributing to the effect estimate of 0.70 in the comparison of verbal face-to-face feedback to no feedback after excluding studies with a high risk of bias. Hence face-to-face feedback may result in a moderate to large improvement in health professionals' workplace task performance.

References

1. Ahlborg L, Weurlander M, Hedman L, et al. Individualized feedback during simulated laparoscopic training:a mixed methods study. *International Journal of Medical Education* 2015;6:93-100. doi: https://dx.doi.org/10.5116/ijme.55a2.218b
2. O'Connor A, Schwaitzberg SD, Cao CG. How much feedback is necessary for learning to suture? *Surgical Endoscopy* 2008;22(7):1614-9. doi: https://dx.doi.org/10.1007/s00464-007-9645-6
3. Pavo N, Goliasch G, Nierscher FJ, et al. Short structured feedback training is equivalent to a mechanical feedback device in two-rescuer BLS: a randomised simulation study. *Scandinavian Journal of Trauma, Resuscitation & Emergency Medicine* 2016;24:70. doi: https://dx.doi.org/10.1186/s13049-016-0265-9
4. Xeroulis GJ, Park J, Moulton CA, et al. Teaching suturing and knot-tying skills to medical students: a randomized controlled study comparing computer-based video instruction and (concurrent and summary) expert feedback. *Surgery* 2007;141(4):442-9. doi: https://dx.doi.org/10.1016/j.surg.2006.09.012

5. Olms C, Jakstat HA, Haak R. The Implementation of Elaborative Feedback for Qualitative Improvement of Shade Matching-A Randomized Study. *Journal of Esthetic & Restorative Dentistry* 2016;28(5):277-86. doi: 10.1111/jerd.12231
6. Bonrath EM, Dedy NJ, Gordon LE, et al. Comprehensive Surgical Coaching Enhances Surgical Skill in the Operating Room: A Randomized Controlled Trial. *Annals of Surgery* 2015;262(2):205-12. doi: https://dx.doi.org/10.1097/SLA.0000000000001214
7. Boyle E, O'Keeffe DA, Naughton PA, et al. The importance of expert feedback during endovascular simulator training. *Journal of Vascular Surgery* 2011;54(1):240-48.e1. doi: https://dx.doi.org/10.1016/j.jvs.2011.01.058
8. Soucisse ML, Boulva K, Sideris L, et al. Video Coaching as an Efficient Teaching Method for Surgical Residents-A Randomized Controlled Trial. *Journal of Surgical Education* 2017;74(2):365-71. doi: https://dx.doi.org/10.1016/j.jsurg.2016.09.002
9. Kroft J, Ordon M, Po L, et al. Preoperative Practice Paired With Instructor Feedback May Not Improve Obstetrics-Gynecology Residents' Operative Performance. *Journal of Graduate Medical Education* 2017;9(2):190-94. doi: https://dx.doi.org/10.4300/JGME-D-16-00238.1
10. Vafaei A, Heidari K, Hosseini MA, et al. Role of feedback during evaluation in improving emergency medicine residents' skills; an experimental study. *Emergency* 2017;5 (1) (no pagination)(e28)
11. Skeff KM. Evaluation of a method for improving the teaching performance of attending physicians. *American Journal of Medicine* 1983;75(3):465-70.
12. Ivers N, Jamtvedt G, Flottorp S, et al. Audit and feedback: effects on professional practice and healthcare outcomes. *Cochrane Database Syst Rev* 2012(6):CD000259. doi: https://dx.doi.org/10.1002/14651858.CD000259.pub3
13. Sterne JAC, Savovic J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. *British Medical Journal* 2019;366:l4898. doi: 10.1136/bmj.l4898 [published Online First: 2019/08/30]
14. Guyatt G, Oxman AD, Sultan S, et al. GRADE guidelines: 11. Making an overall rating of confidence in effect estimates for a single outcome and for all outcomes. *Journal of Clinical Epidemiology* 2013;66(2):151-57. doi: 10.1016/j.jclinepi.2012.01.006