Adv in Health Sci Educ (2012) 17:27–37
DOI 10.1007/s10459-011-9291-6

# Challenging medical students with an interim assessment: a positive effect on formal examination score in a randomized controlled study

**Marleen Olde Bekkink · Rogier Donders · Goos N. P. van Muijen · Dirk J. Ruiter**

Received: 15 November 2010 / Accepted: 15 March 2011 / Published online: 27 March 2011


**Abstract** Until now, positive effects of assessment at a medical curriculum level have not been demonstrated. This study was performed to determine whether an interim assessment, taken during a small group work session of an ongoing biomedical course, results in students' increased performance at the formal course examination. A randomized controlled trial was set up, with an interim assessment without explicit feedback as intervention. It was performed during a regular biomedical Bachelor course of 4 weeks on General Pathology at the Radboud University Nijmegen Medical Centre. Participants were 326 medical and 91 biomedical science students divided into three study arms: arm Intervention-1 (I-1) receiving one interim assessment; arm I-2 receiving two interim assessments, and control arm C, receiving no interim assessment. The study arms were stratified for gender and study discipline. The interim assessment consisted of seven multiple-choice questions on tumour pathology. Main outcome measures were overall score of the formal examination (scale 1–10), and the subscore of the questions on tumour pathology (scale 1–10). We found that students who underwent an interim assessment (arm I) had a 0.29-point (scale 1–10) higher score on the formal examination than the control group ($p = 0.037$). For the questions in the formal examination on tumour pathology the score amounted to 0.47 points higher ($p = 0.007$), whereas it was 0.17 points higher for the questions on topics related to the previous 3 weeks. No differences in formal examination score were found between arms I-1 and I-2 ($p = 0.817$). These findings suggest that an interim assessment during a small group work session in a randomized study setting stimulates students to increase their formal examination score.

M. Olde Bekkink · D. J. Ruiter (✉)
109 Department of Anatomy, Radboud University Nijmegen Medical Centre, P.O. Box 9101, 6500 HB Nijmegen, The Netherlands
e-mail: D.Ruiter@pathol.umcn.nl

R. Donders
Department of Epidemiology, Biostatistics and Health Technology Assessment, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands

M. Olde Bekkink · G. N. P. van Muijen · D. J. Ruiter
Department of Pathology, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands

**Keywords** Interim assessment · Increased examination score · Medical education · Student's performance · Test enhanced learning

## Introduction

Doctors' clinical reasoning skills depend highly on a relevant knowledge base (van der Vleuten and Newble 1995). Becoming an excellent doctor starts at medical school. In order to promote excellence in medical teaching and learning, it is necessary to find out how teaching affects learning (Ramani 2006). One could wonder whether our medical students are being optimally stimulated. Is the active learning of students sufficient, or can they be stimulated to perform even better? For this purpose, objective information on learning efficacy is needed. Assessment of learning efficacy currently involves an integrated approach of formative and summative assessments, and regular evaluation of competences, that are recorded in a student portfolio (Driessen et al. 2005; Epstein 2007). Recently, the role of interim assessments as a third type of assessment in a comprehensive assessment system of US school districts was described, that: (1) evaluates students' knowledge and skills relative to a specific set of academic goals, typically within a limited time frame; and (2) are designed to inform decisions at the classroom level and beyond (Perie et al. 2007). Interim assessments contain both formative and summative assessment features, but unlike true formative assessments, the results of interim assessments can be meaningfully aggregated and reported at a broader level. An interim assessment reflects the level of the students' knowledge and skills, but unlike summative assessments, does not have strict consequences, i.e. pass or fail the assessment. Perie et al. see three different general classes of purposes for interim assessments: instructional; evaluative; and predictive (Perie et al. 2007). All three assessment purposes potentially provide useful information for both students and faculty, and they may also allow further scientific elaboration.

An important goal of assessment is to optimize the capabilities of all learners and practitioners by providing motivation and direction for future learning (Epstein 2007). Assessment also drives students' learning behaviour (Cohen-Schotanus 1999; Frederiksen 1984; van der Vleuten and Schuwirth 2005). Assessment and learning are related to varying degrees, although the specific dynamics are not yet fully understood (Boulet 2008; Handfield-Jones et al. 2002). Apart from obtaining useful information from assessments, it is supposed that assessing drives, and may help learning, the so-called "testing effect" (Newble and Jaeger 1983). Karpicke and Roediger elegantly demonstrated the critical importance of retrieval practice in consolidating learning a foreign language by university students using repeated testing (Karpicke and Roediger 2008). A similar effect was demonstrated by the same authors in two experiments giving students one or three immediate recall tests without feedback (Roediger and Karpicke 2006b). A positive effect of test-driven learning was recently demonstrated in a didactic conference for paediatric and emergency medical residents (Larsen et al. 2009). Thus, assessment can be viewed as an educational tool that provides useful information for both students and faculty (Krupat and Dienstag 2009).

Until now, according to Norman et al. positive effects of assessment at a medical curriculum level have not been demonstrated (Norman et al. 2010). Does interim assessment also improve student performance in a non-laboratory undergraduate medical education setting? If so, we hypothesized that interim testing of the medical students results in a higher formal examination score. Here the interim assessment is used as a didactic

instrument. Medical education uses a variety of settings and formats. Identification of which educational setting lends itself to test-enhanced learning is to be investigated (Larsen et al. 2008). We assumed that the best learning environment to administer the interim assessment is a small group work session, as it is considered to substantially contribute to meaningful learning (Michael 2006). Furthermore, we were interested if we could demonstrate an additional value of two interim assessments instead of one assessment. For this purpose, we set up a prospective randomized study comparing two different arms of small groups. In the intervention arm (I) an interim assessment was provided prior to the formal course examination, in the control arm (C) no interim assessment was provided. The intervention arm was further subdivided into two arms: one arm with one interim assessment (I-1) and the other arm with two interim assessments (I-2). The current study shows that an interim assessment in a randomized study setting is found to stimulate students to increase their formal examination score.

## Methods

### Participants and setting

The study was conducted with biomedical students at the Radboud University Nijmegen Medical Centre, consisting of 326 medical and 91 biomedical science students, who undertook a second-year Bachelor course on General Pathology. The female to male ratio of students was 3:1. The Radboud University Nijmegen Medical Centre provides a learner outcome-oriented curriculum in which each course consists of 4 weeks. The subsequent topics of the course on General Pathology were: (1) Principles of diagnosis and cellular damage; (2) Inflammation and repair; (3) Circulatory disorders; and (4) Tumour pathology (pathogenesis and progression). Each topic had a consistent sequence of educational activities: lecture; task-driven directed self-study in preparation for the subsequent small group work; small group work (obligatory); practical course (obligatory); interactive lecture; and non-directed self-study (see Fig. 1). The formal examination of all topics took place on the final day of the course. For the interim assessment, the small group work session on tumour pathology: "The pathogenesis of uterine cervical carcinoma" was selected.

### Ethical considerations

Formal written permission to execute the study was obtained from the course coordinator. As there is no access to a formal ethical approval process for medical education research in the Netherlands, information about the treatment of the students is provided. This concerns the possible risks for the students, the equitability of the selection, the guarantee of privacy and confidentiality, the procedure on informed consent, and the possible safeguards to protect vulnerable populations (Eva 2009; Kanter 2009). In our opinion, participation in the interim assessments bore no possible risk to students. The assignment of the students to the small groups and the assignment of the groups to one of the three arms of the study was random. The privacy of the students was guarded by the study coordinator. For the study, the examination scores were linked to a student number and the identity of the students was not disclosed. The students were adequately informed of the purpose of the interim assessment and consent was obtained. We were not aware of any vulnerable population among the students that would have required safeguards. When developing the current

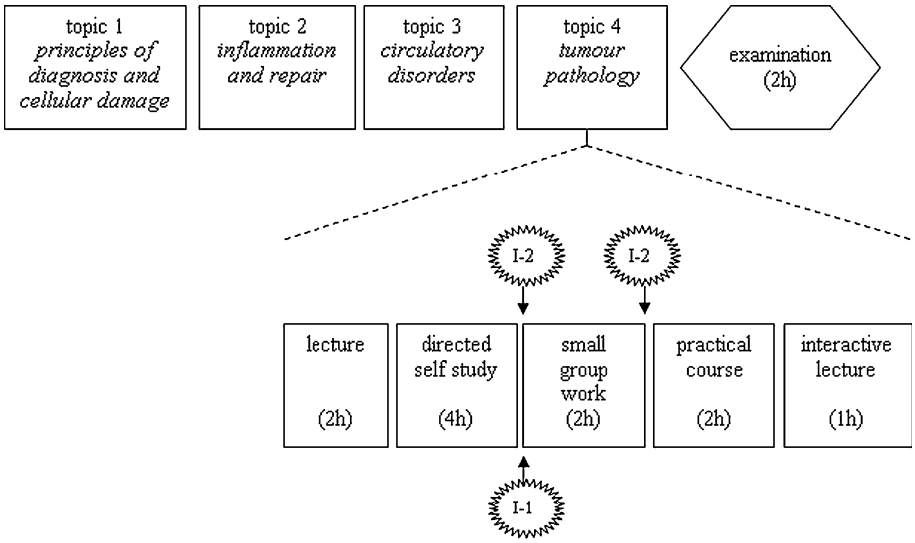


**Fig. 1** Topic structure. Time of administration of a single interim assessment (study arm I-1) and double interim assessments (study arm I-2) in relation to topic structure. The time scheduled is indicated between *brackets* for each educational component

study, the ethical principles of the World Medical Association Declaration of Helsinki were taken into account.

## Intervention

An interim assessment consisted of seven multiple-choice questions with a maximum of four alternative answers on the topic of tumour pathology. A time of 10 min was allotted to each interim assessment. The questions were derived from a bank of 80 multiple-choice questions on tumour pathology formulated by one of the authors (DR), who is an expert in tumour pathology, and were validated by two independent pathologists, two independent medical educationalists, and a master medical student (MOB).

The formal examination consisted of 15 multiple-choice questions and one open question relating to tumour pathology and seven open questions on the other topics. Both the multiple-choice questions of the interim assessments and the formal examination were derived from the aforementioned bank of questions. The two interim assessments and the formal examination were composed of different multiple-choice questions, but the content and the level of the questions were similar.

## Randomization

Participants were randomized in three arms of equal numbers of small work groups. Allocation of intervention occurred on the level of the small work groups. The randomization was stratified for gender and study discipline, since these may influence learning behaviour and learning efficacy (Kusurkar et al. 2009). In arm I-1, students underwent an interim assessment once, i.e. at the end of the small group work session; in arm I-2, students underwent an interim assessment twice, i.e. at the beginning and at the end of the small group work session; and in arm C, students did not undergo an interim assessment (see Fig. 2).

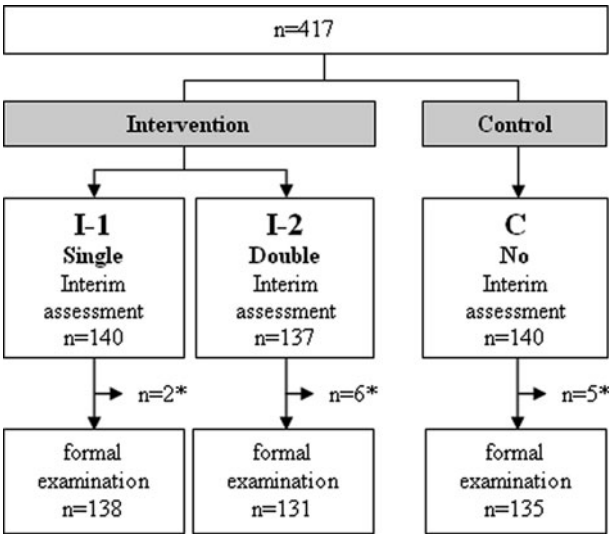


**Fig. 2** Flow chart. Study design including two intervention groups (I-1 and I-2) and one control group (C). *Number of students excluded, because they did not participate in the formal examination (n = 13)

## Procedure

Students in the intervention arm were informed about the interim assessment at the small group work session. Tutors explained to the students immediately before the interim assessment that it was an investigation to inform the faculty on the learning outcome of the students during the small group work. Participation in the interim assessment was on a voluntary basis, and students could stop taking the assessment at any time. They were assured that the result of the interim assessment would not be taken into account for determining the score of the formal course examination. The participation rate was 100%. Students and tutors were not informed of the content of the questions of the interim assessment. The tutors were present at the beginning of the small group work session including the interim assessment, and during the second hour of the small group work session including the other interim assessment. Five different tutors guided the small group work sessions. Each tutor guided both intervention and control groups. No explicit feedback on the results was given to the students. The formal examination took place 3 days following the interim assessments.

## Outcome measures

The main outcome measures were overall score of the formal examination, and the subscore of the open and multiple-choice questions on tumour pathology. Both outcome measures were presented on a scale from 1 to a maximum of 10 points. A subgroup analysis of gender and discipline was performed. The interim assessment is intended as a didactic instrument, not a predictive instrument, therefore the scores of the interim assessment were not compared to that of the formal examination.

### Statistical analysis

Linear mixed models were used in order to account for the dependence caused by clustering of the students into small groups. The small group was used as a random factor. Analysis was performed according to the intention-to-treat principle. After the primary analysis, a subgroup analysis was performed according to gender and discipline.

## Results

### Main results

Students who underwent an interim assessment once or twice (arms I-1 and I-2, respectively) showed a 0.29 point (scale 1–10) higher overall score on the formal examination than the control group C ($p = 0.037$). For the questions in the formal examination related to the topic of tumour pathology, the score amounted to 0.47 points higher ($p = 0.007$), whereas it was 0.17 points higher for the questions of the other topics on general pathology. Accompanying effect scores and standard deviations are reported in Table 1. Results of the mixed model analysis are reported in Table 2. No differences in formal examination score were found between arms I-1 and I-2 (Table 3).

No student refused to participate. Students who undertook the interim assessment, but did not undertake the examination, were excluded (n = 13). A total of 404 students were included in the analysis. There was no significant difference in dropouts between the three study arms.

### Subgroup analysis

Female students scored significantly higher on the formal examination compared with the male students (0.65 points, $p < 0.001$). Medical students scored 0.65 points higher than biomedical science students ($p < 0.001$). There was no difference in progress imposed by the interim assessment between these subgroups.

## Discussion

### Main findings

An interim assessment during a small group work session in a randomized controlled trial setting was able to increase students' formal examination score. This effect was similar for the students who took the interim assessment either once or twice. The increase in the score amounted to almost 0.5 points on a scale of 1–10 for those questions in the formal

**Table 1** Outcome measures (scale 1–10) including standard deviations and effect sizes

| Study arm | Formal examination score (SD) | Subscore on tumour pathology (SD) |
|---|---|---|
| Intervention | 6.27 (1.19) | 6.34 (1.50) |
| Control | 5.98 (1.25) | 5.87 (1.51) |
| Effect size | 0.24 | 0.31 |

**Table 2** Results of the mixed model analysis

| Source | Numerator df | Denominator df | F | Significance |
|---|---|---|---|---|
| a. Type III Tests of fixed effects, dependent variable: formal examination score | | | | |
| Intercept | 1 | 27.235 | 5,906.763 | 0.000 |
| Intervention | 1 | 24.325 | 4.851 | 0.037 |
| Gender | 1 | 399.947 | 27.381 | 0.000 |
| Discipline | 1 | 25.620 | 18.454 | 0.000 |
| b. Type III Tests of fixed effects, dependent variable: subscore on tumour pathology | | | | |
| Intercept | 1 | 27.524 | 3,948.371 | 0.000 |
| Intervention | 1 | 24. 494 | 8.513 | 0.007 |
| Gender | 1 | 399.996 | 17.832 | 0.000 |
| Discipline | 1 | 25.846 | 16.839 | 0.000 |

**Table 3** Results formal examination per intervention arm

| Study arm | Formal examination score (scale 1–10) |
|---|---|
| Intervention-1 | 6.28 (6.40[a]) |
| Intervention-2 | 6.25 (6.27[a]) |

[a] Subscore on tumour pathology

examination that were related to the questions in the interim assessment. There was no difference in progress imposed by the interim assessment between gender or discipline.

## Strengths

The study design, a prospective randomized controlled trial with stratification for gender and discipline can be considered to be robust, because selection bias, information bias and confounding bias are highly unlikely. The primary outcome of the study, i.e. the score of the formal examination, is unequivocal. The data were subjected to a linear mixed-model analysis in order to account for the dependence caused by clustering of the students in small work groups. The multiple-choice questions in the interim assessment and formal examination were validated both on medical content and educational quality. Based on these considerations, the results appear consistent.

The control group was not engaged in an alternative interim assessment, as this would distract from the small group work. The students in the control group could spend time discussing the topic of the small workgroup, when the intervention groups received the interim assessment. Therefore, total exposure time to the subject matter was equal for the intervention and the control groups.

The study setting was directly related to educational practice, i.e. during an ongoing regular biomedical Bachelor course and it did not interfere with educational activities. The tutors were blinded to the content of the interim assessment. All tutors guided at least one student group from each of the three study arms. Both students and tutors accepted the interim assessment well and perceived it as a natural component of the small group work session. Based on regular evaluations, the course on General Pathology is highly appreciated by the students and the faculty, and can be considered to use current best practice. We therefore feel that the study is representative of current best educational practice.

## Limitations

The generalizability of our findings is currently limited. This study presents only a single study in a single curriculum. To increase the level of evidence and to investigate a broader application of the interim assessment, more similar studies are needed.

We were not able to demonstrate an additional learning effect of a second interim assessment in the current study. This might be caused by the length of the interval between de two interim assessments, as will be discussed later.

If our results, that participation in an interim assessment prior to a formal examination increases the score of the formal examination, are confirmed by other studies, this would mean that the students in the interim assessment arms were at an advantage over the students in the control group. Therefore, in future studies, the control group should also be subject to an interim assessment, using cross-over study designs, for example.

Thirteen students (3.1%) could not be included in our analysis, because they did not take part in the formal examination. Among the dropouts the male: female ratio was 5:8 (overall ratio: 1:2), the biomedical: medical ratio was 4:9 (overall ratio: 1:4). The dropouts were distributed equally over the three study arms; therefore it is unlikely this will have affected our results.

## Interpretation of the main findings

As the students were not aware of our study hypothesis, i.e. that participating in an interim assessment would lead to a higher formal examination score, we assume that they were stimulated or even challenged by the interim assessment, as such. By doing so, they probably were engaged in retrieval practice in consolidating learning as a manifestation of the testing effect (Karpicke and Roediger 2008). The underlying mechanisms of this effect may include: (1) enhanced *motivation* of the learners; (2) *directing* them to focus on relevant issues; and (3) giving them an opportunity to *train* for the formal course examination (Larsen et al. 2008). Although the positive effect on the formal examination was relatively small, we feel that it has educational relevance because it could have had a clear influence on the summative exam, i.e. pass or fail. In addition, it demonstrates that students in an ongoing curriculum (i.e. a realistic setting) can be stimulated by an interim assessment to perform better.

The fact that the positive effect on the formal examination score was not different using either one or two interim assessments indicates that a second interim assessment taken within a short time interval (i.e. less than 2 h) following the first interim assessment has no added value on the learning effect. Therefore, it is likely that such an additional effect requires a longer timeframe in between assessments. Karpicke and Roediger demonstrated increased benefits of repeated testing when tests are distributed over time (Karpicke and Roediger 2007). Another factor may be feedback, as it seems a prerequisite for the added value of multiple assessments (Larsen et al. 2008), as will be discussed later.

## Comparison with other studies

An interim assessment is a relatively new educational tool that has recently been developed in the context of secondary schools in the USA (Perie et al. 2007). Repeated testing during a course, that leads to better retention of information, could be considered as a series of interim assessments. Poljicanin et al. demonstrated a positive effect of daily mini quizzes

on students' performance in an anatomy course (Poljicanin et al. 2009). They conducted a total of 34 quizzes during a whole academic year; whereas in our study, we provided only one or two assessments in a 4-week course. It is to be investigated how many assessments per timeframe would gain an optimal increase in performance, without interfering with the regular course programme. Karpicke and Roediger demonstrated that repeated testing leads to better long-term recall in comparison with single testing (Karpicke and Roediger 2008). In the current study, we were not able to demonstrate this result, as there was no significant difference between the intervention groups taking one or two interim assessments. As stated before, this can be explained by the fact that both tests were applied in the same small group work session, with only 2 h in between. It would be interesting to investigate whether the timing of the interim assessment, i.e. either at the beginning or at the end of the small group, would matter in this respect.

Larsen and colleagues recently described improvement of long-term retention by medical residents following repeated testing in a real-life educational setting (Larsen et al. 2009). In contrast to our study, the testing was followed by feedback, and the findings were measured at a final recall interval of 6 months. Our findings suggest that even without such feedback, retention of information, as measured by the formal examination score, occurs. It is conceivable that the increase of the score might have been higher if we would have given feedback as indicated by the literature (Larsen et al. 2008; Roediger and Karpicke 2006a; Wood 2009). For the sake of clarity of the study design, we chose not to include explicit feedback in this study, but we have included it in a follow-up study using a cross-over design. In this new study, we have carefully considered the nature, source and timing of feedback, as suggested by Veloski et al. (2006).

## Conclusions

An interim assessment during a small group work session is found to stimulate students to learn better and to increase their score of the formal examination. The current study supports the efficacy of the testing effect in an ongoing medical curriculum and the view that assessment can be seen as an educational tool (Krupat and Dienstag 2009). An interim assessment may enrich the repertoire of formats of small group work as suggested by Michael, in order to further increase meaningful learning (Michael 2006). It also implies that in our current educational best practice, students still can be challenged to promote excellence in medical education. Further randomized controlled studies assessing the frequency of testing and the addition of feedback are needed to optimize the test-enhanced increase in student performance in a realistic educational setting.

**Acknowledgments** The authors are grateful to Doctor Peter de Wilde, Radboud University Nijmegen Medical Centre, for his contribution to the pilot version of this study. We would also like to thank Ms Xandra Smits, IOWO Consultancy and Institutional Research in Higher Education, for her skilful assistance in data processing, and Professor Pieter de Vries Robbé, Radboud University Nijmegen Medical Centre, for giving valuable comments. Furthermore we would like to thank Professor Piet Slootweg, Doctor Arnold Thoben, Doctor Marc Vorstenbosch and Doctor Rob de Waal, Radboud University Nijmegen Medical Centre, for validating the questions of the interim assessment. This research was funded by the Radboud University Nijmegen Medical Centre.

**Conflict of interest** None.

**Ethical approval** Ethical considerations are discussed as a separate paragraph in the "Methods" section.

## References

Boulet, J. (2008). Teaching to test or testing to teach? *Medical Education, 42*, 952–953.

Cohen-Schotanus, J. (1999). Student assessment and examination rules. *Medical Teacher, 21*, 318–321.

Driessen, E., van der Vleuten, C., Schuwirth, L., van Tartwijk, J., & Vermunt, J. (2005). The use of qualitative research criteria for portfolio assessment as an alternative to reliability evaluation: A case study. *Medical Education, 39*, 214–220.

Epstein, R. M. (2007). Assessment in medical education. *The New England Journal of Medicine, 356*, 387–396.

Eva, K. W. (2009). Research ethics requirements for Medical Education. *Medical Education, 43*, 194–195.

Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. *American Psychologist, 39*, 193–202.

Handfield-Jones, R. S., Mann, K. V., Challis, M. E., Hobma, S. O., Klass, D. J., McManus, I. C., et al. (2002). Linking assessment to learning: A new route to quality assurance in medical practice. *Medical Education, 36*, 949–958.

Kanter, S. L. (2009). Ethical approval for studies involving human participants: Academic medicine's new policy. *Academic Medicine, 84*, 149–150.

Karpicke, J. D., & Roediger, H. L. III (2007). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 33*, 704–719.

Karpicke, J. D., & Roediger, H. L. III (2008). The critical importance of retrieval for learning. *Science, 319*, 966–968.

Krupat, E., & Dienstag, J. L. (2009). Commentary: Assessment is an educational tool. *Academic Medicine, 84*, 548–550.

Kusurkar, R., Kruitwagen, C., Ten Cate, O., Croiset, G. (2009). Effects of age, gender and educational background on strength of motivation for medical school. *Advances in Health Sciences Education: Theory and Practice*. doi:10.1007/s10459-010-9253-4.

Larsen, D. P., Butler, A. C., & Roediger, H. L. III (2008). Test-enhanced learning in medical education. *Medical Education, 42*, 959–966.

Larsen, D. P., Butler, A. C., & Roediger, H. L. III (2009). Repeated testing improves long-term retention relative to repeated study: A randomised controlled trial. *Medical Education, 43*, 1174–1181.

Michael, J. (2006). Where's the evidence that active learning works? *Advances in Physiology Education, 30*, 159–167.

Newble, D. I., & Jaeger, K. (1983). The effect of assessments and examinations on the learning of medical students. *Medical Education, 17*, 165–171.

Norman, G., Neville, A., Blake, J. M., & Mueller, B. (2010). Assessment steers learning down the right road: Impact of progress testing on licensing examination performance. *Medical Teacher, 32*, 496–499.

Perie, M., Marion, S., Gong, B., Wurtzel, J. (2007). The role of interim assessments in a comprehensive assessment system: A policy brief. http://inpathways.net/role-interim.pdf.

Poljicanin, A., Caric, A., Vilovic, K., Kosta, V., Marinovic Guic, M., Aljinovic, J., et al. (2009). Daily mini quizzes as means for improving student performance in anatomy course. *Croatian Medical Journal, 50*, 55–60.

Ramani, S. (2006). Twelve tips to promote excellence in medical teaching. *Medical Teacher, 28*, 19–23.

Roediger, H. L., & Karpicke, J. D. (2006a). The power of testing memory. Basic research an implications for educational practice. *Perspectives on Psychological Science, 1*, 181–206.

Roediger, H. L., & Karpicke, J. D. (2006b). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*, 249–255.

van der Vleuten, C. P., & Newble, D. I. (1995). How can we test clinical reasoning? *Lancet, 345*, 1032–1034.

van der Vleuten, C. P., & Schuwirth, L. W. (2005). Assessing professional competence: From methods to programmes. *Medical Education, 39*, 309–317.

Veloski, J., Boex, J. R., Grasberger, M. J., Evans, A., & Wolfson, D. B. (2006). Systematic review of the literature on assessment, feedback and physicians' clinical performance: BEME Guide No. 7. *Medical Teacher, 28*, 117–128.

Wood, T. (2009). Assessment not only drives learning, it may also help learning. *Medical Education, 43*, 5–6.

World Medical Association Declaration of Helsinki: Ethical principles for medical research involving human subjects, adopted in 1964, readopted and revised in 2008, http://www.wma.net/en/30publications/10policies/b3/17c.pdf.