



RESEARCH ARTICLE



Using longitudinal progress test data to determine the effect size of learning in undergraduate medical education – a retrospective, single-center, mixed model analysis of progress testing results

Dennis Görlich ^a and Hendrik Friederichs ^b

^aInstitute of Biostatistics and Clinical Research, University of Münster, Germany; ^bMedical Education Research Group, Medical School OWL, Bielefeld University, Bielefeld, Germany

ABSTRACT

Medical education research focuses on the development of efficient learning methods promoting the acquisition of student's knowledge and competencies. Evaluation of any modification of educational approaches needs to be evaluated accordingly and a reliable effect size needs to be reached. Our aim is to provide a methodological basis to calculate effect sizes from longitudinal progress test data that can be used as reference values in further research. We used longitudinally collected progress test data and evaluated the increasing knowledge of medical students from the first to the fifth academic year. Students were asked to participate in the progress test, which consists of 200 multiple-choice questions in single best answer format with an additional 'don't know' option. All available individual test scores of all progress tests ($n = 10$) administered between April 2012 and October 2017 were analyzed. Due to the large amount of missing test results, e.g., from students at the beginning of their studies, a linear mixed model was fitted to include all collected data. In total, we analyzed 6324 test scores provided by 2587 medical students. Mean score for medical knowledge (% correct answers) increases from 16.6% (SD: 10.8%) to 51.0% (SD: 15.7%, overall effects size using linear mixed models $d = 1.55$). Medical students showed a learning effect of $d = 0.54$ (total gain: 6.9%) between the 1st and 2nd, $d = 0.88$ (total gain: 12.0%) between the 2nd and 3rd, $d = 0.60$ (total gain: 7.9%) between the 3rd and 4th and $d = 0.58$ (total gain: 7.9%) between the 4th and 5th study year. We demonstrated that incomplete data from longitudinally collected progress tests can be used to acquire reliable effect size estimates. The demonstrated effects size between $d = 0.53$ –0.9 by study year may help researchers to design studies in medical education.

ARTICLE HISTORY

Received 5 March 2021
Revised 10 June 2021
Accepted 20 August 2021

KEYWORDS



Medical education;
undergraduate; progress
testing; learning; learning
curve; students; medical


Introduction

The teaching and assessing of medical knowledge is a central task of medical faculty, because knowledge has been shown to be a key element of the performance of medical doctors [1,2]. Thus, medical education research focuses on the acquisition of student's knowledge. The statement of D.A. Cook 'If you teach them, they will learn' [3] is widely proven in the research literature. Cook himself argues with data summarized from four separate meta-analyses [4–7], comparing various forms of training (e.g., internet- or simulation-based education) with no intervention in 750 studies in medical education. Additionally, there is a huge amount of research that supports this conclusion for education in general. In his famous meta-synthesis, Hattie analyzed the results of more than 800 meta-analyses of learning in school and found a positive impact on learning in about 95% of all interventions. In consequence, he also states that nearly everything works [8].

But learning is not only a question of if you learn, it is also a question of how much you learn. Therefore, because teaching is very resource-binding (e.g., time resources like the workload of teachers, the learning time of pupils and students and material resources like the material, rooms, etc.), it is important to assess not only the effectiveness of learning but also the efficiency.

Consequently, previous research has tried to establish minimally important effect sizes as reference standards in learning ('effect size' as a number measuring the strength of the relationship between a teaching intervention and the learning of students). In Hattie's work – where learning is primarily defined as the gain in knowledge per year – he has documented effect sizes (in Cohen's d) spanning from -0.34 for 'mobility' over 0.5 for 'reading recovery program' to 1.44 for 'student self-reported grades'. He could show that the average effect size of learning of pupils is 0.39 and concludes that effect sizes of bigger than 0.6 are of 'high influence' [9]. Therefore, Hattie

CONTACT Hendrik Friederichs  friedeh@uni-muenster.de  Medical Education Research Group, Medical School OWL, Bielefeld University, Bielefeld, Germany

 Supplemental data for this article can be accessed [here](#).

© 2021 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

demands to use teaching interventions which have an effect size of 0.4 at minimum. So, if there is enough knowledge about the learning effects in your target group you can define a kind of minimally important difference.

On the other hand, Bloom (1984) draws our attention to the optimum effect size of learning [10]. He emphasizes the role of one-to-one tutoring with effect sizes about 2.0 in comparison to other methods of group instruction (with lower effect sizes). Especially medical students must gain a large amount of knowledge and so medical studies courses are very long-lasting and expensive. In consequence, effective and efficient methods of group instructions with a minimum level of effect size are mandatory. The degree of acquisition of domain-specific knowledge by students is one of the measures of the effectiveness of a medical curriculum [11].

To assess cumulative increases in medical knowledge progress testing is becoming increasingly popular internationally [12]. Introduced independently in the late 1970s at the University of Missouri-Kansas City School of Medicine [13] and at Maastricht University in the Netherlands [14,15] it is now used in medical programs across the world. A published guide of the Association for Medical Education in Europe (AMEE) describes the following key elements of progress testing [16]:

administration to all students in an academic program

- regular intervals of testing throughout the academic program
- sampling of the complete knowledge domain expected of students at the end of their course, regardless of the year level of the student

In Germany, medical schools are offered a progress test from Berlin Charité, in which actually 14 medical schools take part in. Because progress testing can be used to compare curricular changes [17–19] most of them want to monitor the learning of their students in reform curricula in comparison with traditional curricula. Typical examples are a PBL-based curriculum at Berlin Charité or a reform-oriented curriculum at Bochum faculty which both had a parallel traditional track. In particular, stand-alone reform-oriented curricula use this kind of supervision for curricular development. Progress tests are a comprehensive examination of the complete final objectives of the curriculum [14]. Therefore, the German progress test contains approximately 200 items in a multiple-choice format. Students are discouraged from making blind guesses by giving them the option of stating that they do not know the answer. Because not being summative, medical students usually do not prepare for the test which gives the opportunity to measure retrievable and lasting knowledge. Meanwhile progress testing is proven as

a reliable tool [12] and therefore can be used to measure the growth of medical knowledge.

So far comparatively few attempts have been made to measure the ‘real world’ effect size of learning in medical studies with progress test data [17–20]. In these studies, the growth of knowledge was expressed in absolute percentages. But to compare effects across different groups relative effects tend to be substantially more stable than absolute benefits [21]. To calculate comparable and easy-to-handle effect sizes often Cohen’s *d* or the nearly-equivalent Hedges’ *g* from random-effects meta-analysis are used (e.g., see [3,8]). But to estimate the effect size(s) of knowledge gain in cohorts with follow-up statistical methods suitable for repeated measurements have to be applied (*for more details see the method section*). After some changes in the beginning years of offering the progress test at the faculty of Muenster, till summer 2012 we have a stable application of this assessment and now have got data of 4 years. It is the time to take a look at the progress of our medical students in learning. The aim of this paper is to analyze progress test data to get reference values for efficient learning in undergraduate medical education. In a longitudinal, cross-sectional design to study knowledge growth our contributions are:

- calculating effect sizes by study year and for the whole course of studies
- detecting differences in the course of the studies, especially comparing the two major clinical subjects internal medicine and surgery
- approaching to a specified reference value for effects size of learning in undergraduate medical education

Methods

For our study, we used a cohort design to evaluate the increasing knowledge of medical students from the first to the fifth academic years. In Germany, medical school is completed in 6 years, and students enter the program directly from secondary education. The course of study is divided into preclinical (first 2 years) and clinical (last 4 years) sections. In the last year, or ‘practical’ year, students rotate through various hospital departments. In Germany, a National Competence Based Catalogue of Learning Objective for Undergraduate Medical Education (NKLM) came into effect in June 2015 [22]. Many of the competences described in the NKLM now include the acquisition of basic practical skills but – in accordance with common international practice [23] – still the main part of the entire curriculum is based on the transmission of knowledge-based content.

The study was conducted at the medical school of the University of Muenster, Germany. All students who entered the medical faculty of Muenster between

October 2011 and October 2017 were included in the study. Students are asked to participate voluntarily and anonymously in the progress test, which takes place annually for them in the middle of the study year. Due to admission every semester, medical faculty offers the progress test twice a year to assess cumulative increases in medical knowledge of their students. Every student solves five progress tests, so test scores provide cross-sectional and longitudinal data.

The Berlin Progress Test consists of 200 multiple-choice questions in single best answer format. Multiple-choice questions are selected from an item database and matching a blueprint. After being part of a test, questions are not used for 2 years to prevent collection and simple recall of items [24]. As in most other progress tests a 'don't-know' option is given as not all students (especially those in the first 3 years) are expected to cope with all objectives in the test. Students are encouraged to make use of the 'don't know' option in order to provide a more reliable feedback [24,25] and to reduce the measurement error which could result from random guessing. The students are asked to take the test in about 3 hours. In general, the German progress test shows significant correlation with the German National Licensing Exam (criterion validity) [26].

For the individual feedback the student's test score is obtained by negative marking of incorrect answers, whereas choosing a 'don't know' option has no effect on the individual score. In this so-called 'formula scoring' the number of correct minus incorrect answers was used as the test score. In the present study, only the correct answers of the students were counted and expressed as the percentage of all multiple-choice questions. We collected all available individual test scores of all progress tests ($n = 10$) administered between April 2012 and October 2017 and used these to calculate the average test score for each study year.

Statistical methods

Sample size for this retrospective analysis was not determined beforehand. The final data set contains data of all available tests taken in the prechosen time frame.

Due to the anonymity of the test data and general data protection the study cohort cannot be described by any socio-demographic factors.

We report the number of individual students per study year and the number of tests taken.

The data was cleaned before analysis: (i) We omitted test results of students after their final exam or before their regular start of studying. (ii) We also omitted test results marked as 'irregular participations'. These were participations which were interrupted at an early stage or were noticed by irregular patterns in the marking of

results. (iii) If a participant took the test several times within the same study year only the first occurrence was included in the dataset.

For each individual test the following parameters were calculated: percent of correctly answered questions of all asked questions and percent of correctly answered questions of only the answered questions. The latter, thus, can be interpreted as overall knowledge over questions the students were confident to answer, while the former also incorporates 'don't know' choices as wrong answers. The distribution of percentage of correctly answered questions will be shown as boxplots per study year including mean and median. Additionally, we intended to show the relation between the number of answered questions and the percentage of correctly answered questions (test score) in scatterplots per study year.

To estimate the effect size(s) of knowledge gain we applied statistical methods suitable for repeated measurements. Here a (generalized) linear mixed model was fitted using a normal distribution and identity link function. As fixed effect the study year was included (i.e., time). Repeated measurements were modelled by including random intercepts for the individual students. A 1st order autoregressive covariance matrix was chosen for the model.

To assess the effect (i.e., knowledge gain) between two study years we first calculated the least squares estimates in the statistical model (M_{diff}). Then the effect size d_{GLMM} , similar to Cohen's d_z for paired data [27], is given by $d_{GLMM} = \frac{M_{diff}}{SE_{diff} \cdot \sqrt{N}}$, with M_{diff} as least squares estimate, SE_{diff} as standard error of the estimate and N as total sample size, i.e., the number of individual students providing test data either in the first or second time point.

Finally, we compared knowledge in basic medical knowledge (anatomy [with biology], physiology [with physics], biochemistry [with chemistry and molecular biology], medical psychology and medical sociology) vs clinical knowledge and internal medicine vs surgery.

Calculated effect sizes for all individual subjects are reported in the Supplemental Digital Appendix.

Statistical analysis have been conducted using the SAS® Statistical Software (Version 9.4, SAS Inc. Cary, NC, USA.)

Ethical considerations

Students gave their written consent to data collection at the beginning of their studies, covering also all progress tests. Only anonymized data has been included into the analyzed data set. Best practices in data protection and data security has been adhered to. A vote by the institutional review board was not requested.

Results

In total, we collected 6,546 test scores of 2,656 medical students. Uncompleted tests and tests that show irregular answer patterns were excluded, resulting in an analysis set of 6,324 test scores provided by 2,587 students. Collected tests distribute to the study years as follows: 1st year: $N = 1,463$; 2nd: $N = 1,433$; 3rd: $N = 1,245$; 4th: $N = 1,107$; 5th: $N = 1,076$. While 707 students (27%) took the test only once (irrespective of study year), 1,880 students (73%) repeated the test at least once in a different study year. A detailed participation pattern is given in Table S1. 61.85% of included students were female. The mean age was 23.2 years (SD 4.5); 94% of students were younger than 30 years at their first participation in the progress test. Figure 1A presents the distribution of total test scores for the five study years/measurement points.

Total knowledge gain

In total, the mean score for medical knowledge (% correct answers) increases from 16.6% (Standard deviation (SD): 10.8%) to 50.9% (SD: 15.7%) during the examined phase of study. Using the naive approach to estimate Cohen's d our data would show an effect size of $d = 2.47$. Nevertheless, due to the repeated measure structure of the data we cannot estimate an unbiased effect size for this overall effect based on classical measures (e.g., Cohen's d). Thus, in the following paragraphs, we used linear mixed models to estimate a more conservative effect size under consideration of repeated measures (cp. Methods section).

Learning of medical students has an overall effect size of $d_{\text{GLMM}} = 1.55$ (estimate total gain: 34.8% (95%-CI: 34.0–35.8%), linear mixed model, Figure 1B) for the whole course of study. Analyzing the total knowledge gain between sequential measurement points we observe a learning effect of $d_{\text{GLMM}} = 0.54$ (total gain: 6.9%) between the 1st and 2nd study year and $d_{\text{GLMM}} = 0.88$ (total gain: 12.0%) between the 2nd and 3rd study year. Learning between the 3rd and 4th, and the 4th and 5th study year show effects of $d_{\text{GLMM}} = 0.60$ (total gain: 7.9%), and $d_{\text{GLMM}} = 0.58$ (total gain: 7.9%). See Figure 1B.

Along with the total gain in knowledge the number of answered questions also raises from 31.3% answered questions, on average, in the 1st study year to 67.3% answered questions in the 5th study year, indicating a gain in self-confidence with respect to the own knowledge (Figure 1A). Overall, medical students answer comparatively more questions with relatively more correct solutions in the course of studies (Figure 2).

Comparing basic medical knowledge vs. clinical knowledge

We compared test scores for the pooled basic medical subjects (estimating knowledge in basic medicine) and pooled clinical subjects (clinical knowledge), separately. Figure 3A shows the observed test scores. Figure 3B summarizes the knowledge gain results. We can observe that students up to the 4th year have a higher relative basic medical knowledge (as intended by the curriculum) than clinical knowledge. In particular, the absolute difference after the 1st study years is 26.3% (basic knowledge) vs. 14.9% (clinical knowledge). During the

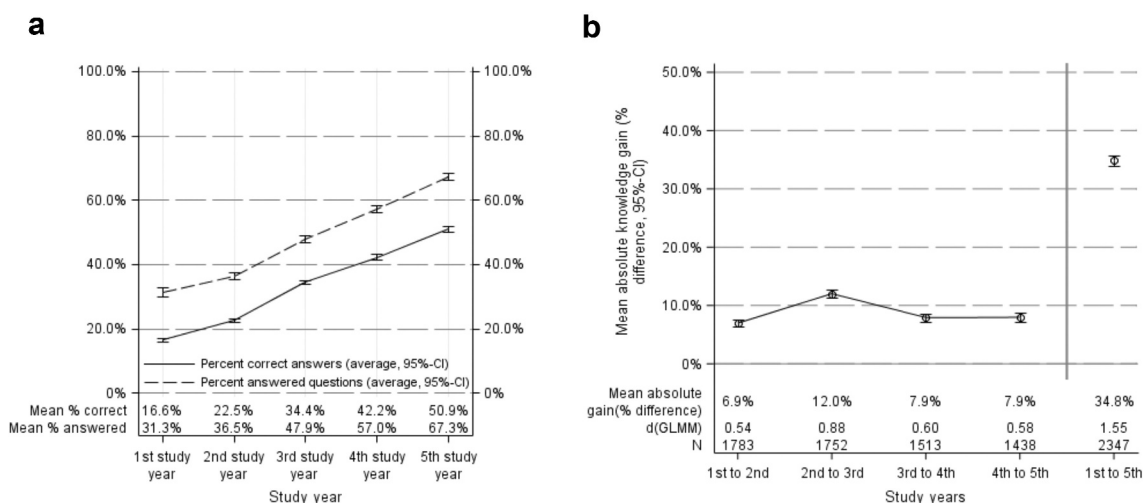


Figure 1. Overview on the collected data. (A) Average percentage of answered questions (dashed line) and the average percentage of correctly answered questions (solid line) over time. Advanced students answer more questions, reflecting the student's overall confidence in their knowledge. Simultaneously, also the percentage of correct answers increases. (B) Absolute knowledge gain (increase in correctly answered questions) between consecutive study years and for the whole curriculum. Absolute gain and calculated effect sizes (linear mixed model) are displayed, respectively.

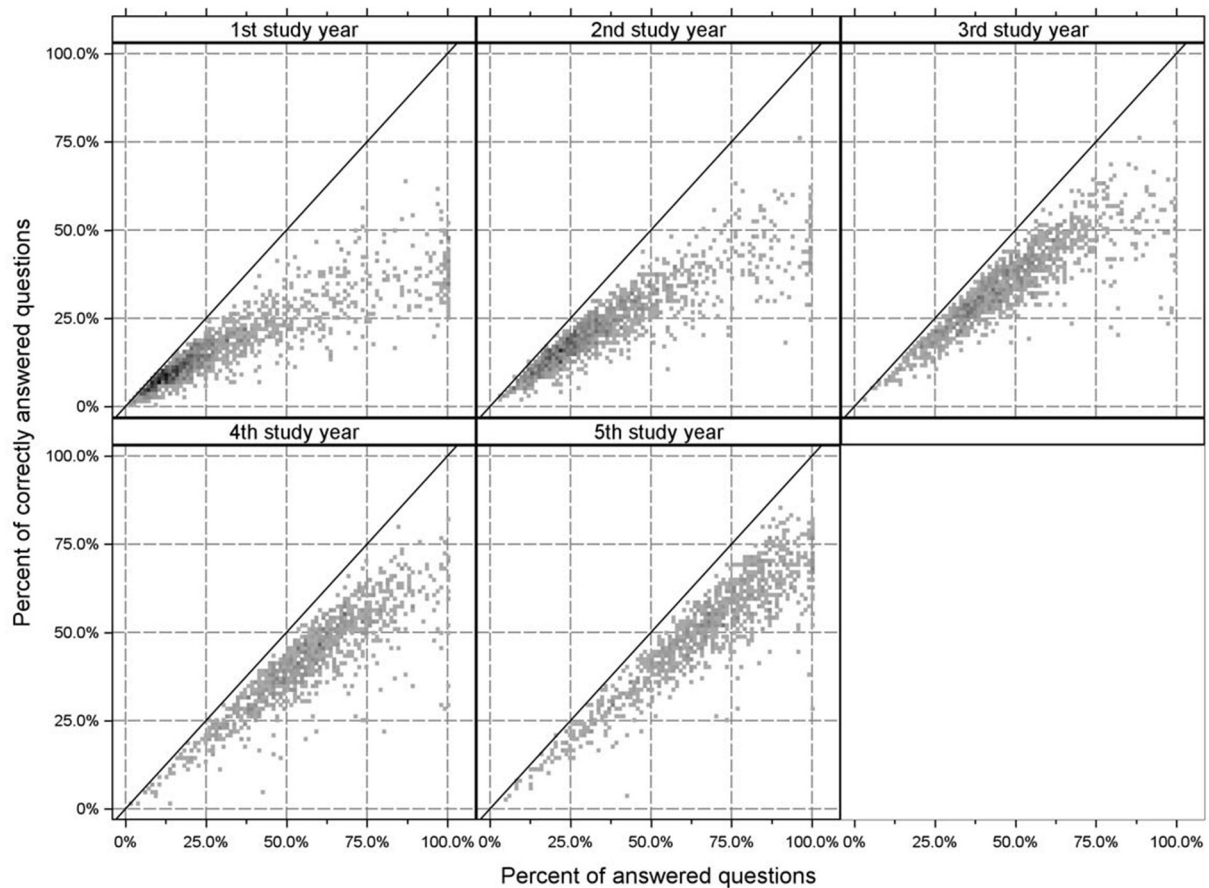


Figure 2. Relation between the number of answered questions and the percentage of correctly answered questions. Data is shown per study year. Darker grey to black color indicates increased density of data points. The straight line marks the maximal reachable percentage of correct answers given the percentage of answered questions. The data shows that students, while getting more confident in answering questions. Over time, students increase their knowledge and give more correct answers with proceeding study years.

course of the first six semesters students increase their basic medical knowledge and then maintain their level of knowledge, after ‘forgetting’ some knowledge between the 3rd and 4th year. On the contrary, clinical knowledge is obtained continuously during the study time. Finally, students reach similar test scores in both domains at the end of their studies. The analysis of effect size between the semesters clearly shows that between the 3rd and 4th year parts of the basic medical knowledge are ‘forgotten’ by the students while the clinical knowledge further increases ($d_{GLMM} = -0.09$ basic knowledge vs. $d_{GLMM} = 0.70$ clinical subjects). The strongest gain in basic medical knowledge is located between the 1st and 2nd year ($d_{GLMM} = 0.77$) – reflecting the courses of basic medical knowledge – while the strongest increase in clinical knowledge is located between the 2nd and 3rd year ($d_{GLMM} = 0.90$).

Comparing internal medicine vs. surgery knowledge

We compared two main subjects, i.e., internal medicine and surgery, directly (Figure 3C shows the average knowledge, Figure 3D the knowledge gain

and effect sizes (d_{GLMM}). Both subjects show a similar knowledge gain profile. Students continuously increase their knowledge with the strongest gain between the 2nd and the 3rd study year (internal medicine: total gain = 14.6%, $d_{GLMM} = 0.87$; surgery: total gain = 11.3%, $d_{GLMM} = 0.44$). Between the 4th and 5th study year almost no new knowledge is gathered, but it seems the level is maintained (internal medicine: total gain = 3.1%, $d_{GLMM} = 0.18$; surgery: total gain = 2.8%, $d_{GLMM} = 0.11$). Over the course of study (1st year to 5th year) internal medicine shows an effect size of $d_{GLMM} = 1.44$ while surgery has an effect size $d_{GLMM} = 1.14$.

Discussion

The average medical knowledge growth curve indicates a steady increase of medical knowledge as hypothesized. Overall, medical students answer comparatively more question with relatively more correct solutions in the course of studies. So, our findings strongly support the view that the progress test is

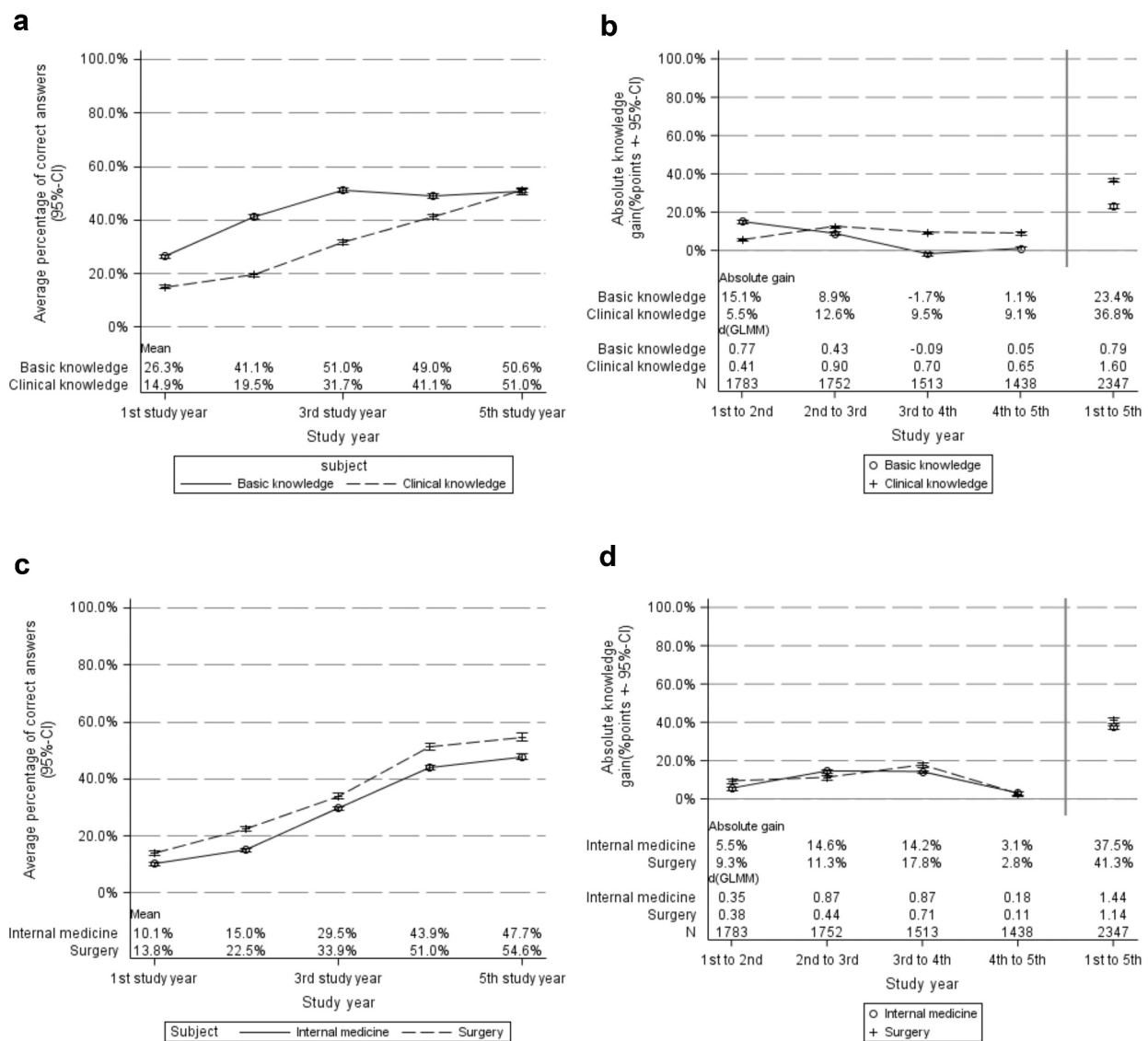


Figure 3. Comparison of major curricular subjects. (A) Percentage of correctly answered questions pooled for basic knowledge and clinical subjects, respectively. (B) Knowledge gain and effect sizes (d_{GLMM}) per study year and the total course of study. (C) and (D) display the same information for comparison of internal medicine and surgery, respectively.

a valid and reliable tool to measure the effect of learning in undergraduate medical education.

Our results suggest that learning of medical students has effects size (d_{GLMM}) between 0.54 and 0.88 by each study year and an effect size of $d_{GLMM} = 1.55$ for the whole course of studies. Growth patterns of medical knowledge in internal medicine and surgery are quite similar. At the end of the entire medical curriculum (except the last or 'practical' year of the German curriculum in which students rotate through various hospital departments), medical students, on average, score 50.9% (SD: 15.7%, IQR: 41.45%–62.24%) of the maximum progress test score.

The size of these effects reflects the achievement potential of medical students and – as a consequence – demands high standards for effective and efficient teaching in medical curricula. In conclusion, it can be noted that teaching interventions in

undergraduate medical education should have an effect size of 0.5 at minimum.

A central question that needs to be addressed in this context is the classification of these effect sizes as high or low. So far comparatively few attempts have been made to give absolute reference values for learning in undergraduate medical education. In general researchers often refer to values given by Cohen. In his influential work on power analysis Cohen strikes the point that – before planning a study – a researcher should ask himself how large he expects the effect in the population. Because it is quite difficult to answer this question Cohen proposes 'ES values to serve as operational definitions of the qualitative adjectives "small", "medium" and "large" as a convention. He clearly states that the definitions are arbitrary and that "they run a risk of being misunderstood' (p. 12). For the comparison of arithmetic

means he proposes a standardization of the raw effect size by dividing the measurement unit of the depended variable by the standard deviation of the respective population, called Cohen's *d*. In general Cohen intended that 'medium ES represent an effect likely to be visible to the naked eye of a careful observer.' and assigned a value $d = 0.5$ (small effect size: $d = 0.2$, large effect size: $d = 0.8$) [28]. In respect to this reference values our findings suggest that learning of undergraduate medical students is medium effective, but has to be ranked context-dependent, too.

As described above, much is known about knowledge acquisition during specific educational interventions. But to date, only a few studies have been published about the growth of medical knowledge during an entire undergraduate medical curriculum. Whereas others are limited to a single discipline or a cluster of disciplines, Verhoeven et al. described the relationship between a problem-based curriculum and the development of students' medical knowledge during the entire training program [11] and found a monotonously increase as a function of training time. In their study, overall knowledge increased from 5% to 41% during the curriculum (mean correct minus incorrect score). Most studies in the field are based on cross-sectional data. For example, Nouns and colleagues [29] presented their progress test results in a comparison of a problem-based curriculum and a traditional approach which showed no differences in gaining knowledge in the follow-ups. Of these data, we could calculate an average effect size of 0.761 (effect sizes calculated with: Ellis PD. Effect size calculators. (<http://www.polyu.edu.hk/mm/sizefaq/calculator/calculator.html>) accessed on 1 February 2018.) per year. Due to quite different sample sizes (e.g., $n = 1,431$ in the third and $n = 529$ in the second study year) and the cross-sectional study design, results of this study have to be handled with caution.

Successive progress test scores reflect the development of medical knowledge throughout the curriculum. The strongest gain in basic medical knowledge is located between the 2nd and 4th semester ($d = 0.81$) and reflects the effectiveness of basic medical knowledge curricula. The strongest increase in clinical knowledge is located between the 4th and 6th semester ($d_{GLMM} = 0.94$) which is at the beginning of clinical knowledge curricula. In our opinion, this data underlines the fundamental role of basic medical knowledge for the acquiring of clinical knowledge [30]. Due to the structure of our curriculum the growth patterns of medical knowledge in internal medicine and surgery are quite similar. Students continuously increase their knowledge with the strongest gain between the 6th and the 8th semester, due to the main lectures and skill trainings located there

(Internal medicine: $d_{GLMM} = 0.97$; Surgery: $d_{GLMM} = 0.78$).

Due to not established measurement points in our study we have not measured the growth of medical knowledge in the 6th year. Raupach et al. have measured the effect size of learning in the last year, or 'practical' year of the German curriculum in which students rotate through various hospital departments [31]. They found an effect size of 0.87 which is in the range of our effect sizes and underlines the importance of teaching practical skills in acquiring medical knowledge.

In total, we found that medical students gain an effect size of 1.62 for the whole course of studies. As mentioned above Bloom (1984) has demonstrated that the achievable effect size of learning [10] is about 2.0 in one-to-one tutoring. The gain of effectiveness in values bigger than two is quite small so that this value represents a kind of optimum in real-life-settings. For example, technology-enhanced simulation training for health professions learners in comparison with no intervention shows pooled effect sizes of 1.20 (95% CI, 1.04–1.35) for knowledge outcomes [5]. Other methods of group instruction are normally found to have lower effect sizes. The value found in the present study may explain the historical development of the long-lasting course of studies in medical education. To gain the expected large amount of knowledge medical students have to learn such a long time.

Strengths and weaknesses of the study

Our findings should be treated as tentative because of several limitations, especially concerning the drop-out students. Due to data integrity, we could not establish a consequent follow-up of all students, so we cannot report the reasons for drop-out. Students could have moved to another university, paused or changed their studies or even gave up studying. Drop-out rates for medical studies in Germany are quite low (less than 15% [32],) but even this rate could have significant influence on the calculated effect sizes. Because more than 85% of students graduate, we decided to use all available data and accept some possible confounding.

Researchers should keep in mind that our research methodology is based on longitudinal assessments and our findings may not be directly applicable to programs where cross-sectional individual subject-based assessments are used at each stage of the program. This latter setting of longitudinal assessments in independent cohorts can be understood as a missing value problem. For each cross-sectional cohort the unobserved time points can be considered missing data. The generalized linear regression model can still be applied to this setting under certain

assumptions. The consequences on effect size estimation, thus, needs to be elucidated in future research.

Conclusions

Growth of medical knowledge at our faculty has an effect size of at least 0.53 per year. In conclusion, we propose to establish a minimally important effect size as a reference standard in undergraduate medical education. As significance criteria (0.05 or 0.01) and desired power (0.80) in educational research are typically constrained by convention, researchers have to think about the expected effect size. Effect sizes are central in research to determine the necessary sample size so that the demonstrated effects size (d_{GLMM}) between 0.53–0.88 by each study year may help researchers to design studies in medical education.

Disclosure statement

The authors declare that they have no competing interests.

Author contributions

HF designed the work and acquired data. DG planned and conducted the statistical data analysis. HF and DG interpreted the results, drafted the manuscript and approved the final version. All authors are accountable for all aspects of the work.

ORCID

Dennis Görlich  <http://orcid.org/0000-0002-2574-9419>
Hendrik Friederichs  <http://orcid.org/0000-0001-9671-5235>

References

- [1] Wenghofer E, Klass D, Abrahamowicz M, et al. Doctor scores on national qualifying examinations predict quality of care in future practice. *Med Educ.* 2009;43(12):1166–1173.
- [2] Glew RH, Ripkey DR, Swanson DB. Relationship between students' performances on the NBME Comprehensive Basic Science Examination and the USMLE Step 1: a longitudinal investigation at one school. *Acad Med.* 1997;72(12):1097–1102.
- [3] Cook DA. If you teach them, they will learn: why medical education needs comparative effectiveness research. *Adv in Health Sci Educ.* 2012;17(3):305–310.
- [4] Cook D, Erwin P, Triola M. Computerized Virtual Patients in Health Professions Education: a Systematic Review and Meta-Analysis. *Acad Med.* 2010;85(10):1589.
- [5] Cook DA, Hatala R, Brydges R, et al. Technology-enhanced simulation for health professions education: a systematic review and meta-analysis. *JAMA.* 2011;306(9):978–988.
- [6] Cook DA, Levinson AJ, Garside S, et al. Internet-Based Learning in the Health Professions: a Meta-analysis. *JAMA.* 2008;300(10):1181–1196.
- [7] McGaghie WC, Issenberg SB, Cohen ER, et al. Does simulation-based medical education with deliberate practice yield better results than traditional clinical education? A meta-analytic comparative review of the evidence. *Acad Med.* 2011;86(6):706–711.
- [8] Hattie J. *Visible Learning.* London: Routledge; 2008.
- [9] Hattie J. *Visible Learning for Teachers.* London: Routledge; 2012.
- [10] Bloom BS. The 2 Sigma Problem: the Search for Methods of Group Instruction as Effective as One-to-One Tutoring. *Educ Researcher.* 1984;13(6):4.
- [11] Verhoeven BH, Verwijnen GM, Scherpbier AJJA, van der Vleuten CPM. Growth of medical knowledge. *Med Educ.* 2002;36(8):711–717.
- [12] Freeman A, van der Vleuten C, Nouns Z. Progress testing internationally. *Med Teach.* 2010;32(6):451–455.
- [13] Arnold L, Willoughby TL. The quarterly profile examination. *Acad Med.* 1990;65(8):515–516.
- [14] van der Vleuten CPM, Verwijnen GM, Wijnen WHFW. Fifteen years of experience with progress testing in a problem-based learning curriculum. *Med Teach.* 2009;18(2):103–109.
- [15] Schuwirth LWT, van der Vleuten CPM. The use of progress testing. *Perspect Med Educ.* 2012;1(1):24–30.
- [16] Wrigley W, van der Vleuten CPM, Freeman A, et al. A systemic framework for the progress test: strengths, constraints and issues: AMEE Guide No. 71. *Med Teach.* 2012;34(9):683–697.
- [17] Bianchi F, Stobbe K, Eva K. Comparing academic performance of medical students in distributed learning sites: the McMaster experience. *Med Teach.* 2008;30(1):67–71.
- [18] Van der Veken J, Valcke M, De Maeseneer J, et al. Impact on knowledge acquisition of the transition from a conventional to an integrated contextual medical curriculum. *Med Educ.* 2009;43(7):704–713.
- [19] Peeraer G, De Winter BY, Muijtjens AMM, et al. Evaluating the effectiveness of curriculum change. Is there a difference between graduating student outcomes from two different curricula? *Med Teach.* 2009;31(3):e64–e68.
- [20] Muijtjens AMM, Schuwirth LWT, Cohen-Schotanus J, et al. van der Vleuten CPM. Benchmarking by cross-institutional comparison of student achievement in a progress test. *Med Educ.* 2008;42(1):82–88.
- [21] Schünemann HJ, Oxman AD, Vist GE. *Chapter 12: interpreting Results and Drawing Conclusions in: Higgins JPT.* In: Green S, editor. *Cochrane Handbook for Systematic Reviews of Interventions.* Chichester (UK): John Wiley & Sons; 2008:403–431
- [22] Fischer MR, Bauer D, Mohn K. NKLM-Projektgruppe. Finally finished! National Competence Based Catalogues of Learning Objectives for Undergraduate Medical Education (NKLM) and Dental Education (NKLZ) ready for trial. *GMS Zeitschrift für Medizinische Ausbildung.* 2015;32(3):Doc35.
- [23] Cooke M, Irby DM, Sullivan W, et al. American medical education 100 years after the Flexner report. *N Engl J Med.* 2006;355(13):1339–1344.
- [24] Nouns ZM, Georg W. Progress testing in German speaking countries. *Med Teach.* 2010;32(6):467–470.
- [25] Schaubert S, Nouns ZM. Using the cumulative deviation method for cross-institutional benchmarking in the Berlin progress test. *Med Teach.* 2010;32(6):471–475.

- [26] Nouns Z, Hanfler S, Brauns K, et al. presented at: AMEE-Conference, 2004 [Edinburgh](#). *Do Progress Tests Predict the Outcome of National Exams*.
- [27] Lakens D. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Front Psychol.* 2013;4. DOI:10.3389/fpsyg.2013.00863.
- [28] Cohen J. *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Erlbaum; 1988.
- [29] Nouns Z, Schauber S, Witt C, et al. Development of knowledge in basic sciences: a comparison of two medical curricula. *Med Educ.* 2012;46(12):1206–1214.
- [30] de Bruin ABH, Schmidt HG, Rikers RMJP. The role of basic science knowledge and clinical knowledge in diagnostic reasoning: a structural equation modeling approach. *Acad Med.* 2005;80(8):765–773.
- [31] Raupach T, Vogel D, Schiekirka S, et al. Increase in medical knowledge during the final year of undergraduate medical education in Germany. *GMS Zeitschrift für Medizinische Ausbildung.* 2013;30(3):Doc33.
- [32] Kadmon G, Resch F, Duelli R, et al. Predictive value of the school-leaving grade and prognosis of different admission groups for academic performance and continuity in the medical course - a longitudinal study. 2014 *GMS Zeitschrift für Medizinische Ausbildung*;31(2):Doc21.