Title

A validity argument for progress testing: Examining the relation between growth trajectories obtained by progress tests and national licensing examinations using a latent growth curve approach.

Authors

Yassin Karay, Dean's Office for Student Affairs, Medical Faculty of the University of Cologne, Germany

Stefan K. Schaubert, Centre for Educational Measurement at the University of Oslo (CEMO) and Centre for Health Sciences Education, University of Oslo, Oslo, Norway

Practice Points

- There are relationships between the individual students' development as obtained by progress tests and performance in high-stakes national licensing examinations.
- We provide evidence for a progress tests suitability to monitor students' growth of knowledge, especially for the clinical phase.
- Analyses of the development of students within a cohort could be used as a tool for early diagnosis of performance issues.

23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46

Abstract

Background

Progress testing is a longitudinal assessment that aims at tracking students' development of knowledge. This approach is used in many medical schools internationally. Although progress tests are longitudinal in nature, and their focus and use of developmental aspects is a key advantage, individual students' learning trajectories themselves play, to date, only a minor role in the use of the information obtained through progress testing.

Methods

We investigate in how far between-person differences in initial levels of performance and within-person rate of growth can be regarded as distinct components of students' development and analyze the extent to which these two components are related to performances on national licensing examinations using a latent growth curve model.

Results

Both, higher initial levels of performances and steepness of growth are positively related to long-term outcomes as measured by performance on national licensing examinations. We interpret these findings as evidence for progress tests' suitability to monitor students' growth of knowledge across the course of medical training

Conclusions

This study indicates that individual development as obtained by formative progress tests is related to performance in high-stakes assessments. Future studies may put more focus on the use of between-persons differences in growth of knowledge.

Introduction

Assessments in educational contexts can be divided into two broad categories: Summative and formative assessments. According to (Wiliam & Black 1996), summative assessments are mostly “designed to judge the extent of students’ learning of the material in a course.” The judgment in a summative assessment may have direct consequences for an individual, if, for instance, the student has passed or failed an exam. By contrast, formative assessments are often characterized as low-stake tests, with an emphasis on providing feedback to students. Usually, the aim of such feedback is to stimulate students’ learning and intrinsic motivation. Hence, the main difference between summative and formative assessment is its intended use. In this respect, Wiliam and Black (1996) emphasize that a classroom interaction (e.g., a teacher asking a question within a lecture) and a formal test (e.g., a national licensing exam) can be seen as two ends of a spectrum of formative assessment. This notion of different uses of assessment is also reflected in the educational literature, where formative assessments have been characterized as *assessments for learning* (Black & Wiliam 2009) while summative assessments have been referred to as *assessments of learning* (Schuwirth & van der Vleuten 2011). Usually, any kind of assessment is both feedback for the student and an indication of the level of mastery students have obtained.

One approach to assessment that blends summative and formative aspects to a high degree is progress testing. Progress testing is a longitudinal assessment that tracks students' learning trajectories throughout their education. With this goal, students are tested repeatedly on the curricular content, typically between 2 to 4 times per academic year (Wrigley et al. 2012). Progress tests do provide the possibility for capturing both individual development and learning trajectories between individuals. Historically, progress testing was a new approach to assessment and was typically implemented in medical curricula that dissolved the strong separation of teaching into several subjects. As teaching and learning became more integrated across subjects and time, so did assessment (Brauer & Ferguson 2015). Nowadays, progress testing is used in many medical schools, in inter-institutional collaborations or for single programs (Aarts et al. 2010; Freeman et al. 2010; Wrigley et al. 2012; Ali et al. 2016; Tio et al. 2016). The basic principle, a focus on an individual student’s learning trajectory, has remained a critical part of the philosophy underlying progress testing – regardless of the way it is implemented.

Despite its conceptual appeal, establishing a progress test in a medical school is an endeavor that takes some effort (Wrigley et al. 2012; Joiner et al. 2016; Neeley et al. 2016; van der Vleuten & Heeneman 2016) and may only be undertaken when its added value can be demonstrated. Indeed, many studies of the use and implementation of progress tests focus on key elements of quality of an assessment and try to highlight the

advantages of this approach. Across scenarios, research indicates that progress tests are usually reliable. Furthermore, several articles indicate that the information obtained can be used to improve the educational environment, as a source of feedback to students, teachers, and deans alike (Coombes et al. 2010; Muijtjens et al. 2010). In addition, more senior students usually outperform more junior medical students. Across the years, students are able to answer more questions – a finding that usually is used as an argument for the validity of a progress test (Dijksterhuis et al. 2009; Nouns et al. 2012; Dijksterhuis et al. 2013). Finally, Blake et al. (1996) reported a standard setting procedure based on individuals' learning trajectories, that is, passing or failing a progress test is in their educational context depended on students' growth of knowledge. Despite this exception, and although progress tests are longitudinal in nature, and their focus and use of developmental aspects is a key advantage, individual students' learning trajectories themselves play, to date, only a minor role in the use of the "rich source of information" (Freeman et al. 2010) obtained through progress testing.

From a mathematical perspective, learning trajectories have at least two basic components. First, any learning begins at a certain level; students usually differ in their prior knowledge on a given subject. When entering medical school, some students may have worked as nurses, emergency assistants, or in other health professions and thus already have relevant knowledge, while others may have a rather unrelated educational background. In the statistical analysis of growth processes, this component is usually referred to as "the intercept." Second, learning evolves over time; students acquire more knowledge but might also have difficulties in retrieving what they learned in the first year. Hence, learning trajectories may be differently steep for different learners – individuals vary in regard to the actual rate of growth across time. This component, the rate of growth, is usually referred to as "the slope." Statistical techniques that allow for modelling these two components at the level of the individual student have both technical and conceptual advantages (Preacher 2008; Leppink 2015; Wimmers & Lee 2015). In the context of progress testing, the most worthwhile feature is that variance components for the intercept and slope can be modelled. Consequently, intercept and slope factors can be related to a third, external variable in order to address the question of how much information the two components carry that would allow us to predict an external criterion.

In this article, we address the question of the extent to which between-person differences in initial levels of performance (intercepts) and within-person rate of growth (slopes) constitute unique sources information on students' development of knowledge. Specifically, we analyze if and to which extent these two components can be disentangled statistically. This is important with respect to the – often rather implicit – assumption that progress tests indeed measure between-person differences in within-person growth. If so, data from progress tests can be legitimately used to formulate hypotheses on students' patterns of growth of knowledge and stress

the possible relation between initial levels of performance (intercepts), growth of knowledge (slopes) and as well as their relation to other criteria.

In order to address the question in how far progress tests – beyond the face-validity of the approach - reflect meaningful information on students within-person development we use a two-step approach. First, we analyze how students' within-person development across the course of study can be modeled and if a model assuming between-person differences in within-person growth can be statistically justified. Second, we analyze the extent to which these two components are related to performances on national licensing examinations. These analyses are carried out using a latent growth curve model (Preacher 2008). Thus, we also contribute to recent calls for using more advanced statistical approaches to investigating research questions in medical education (Leppink 2015). Furthermore, we provide an analysis of long-term development of students' medical knowledge as assessed by low-stakes formative progress tests and their relation to outcomes on a high-stakes national licensing exam. To date, no data have been reported on the relation between initial levels of performance, long-term growth trajectories and outcomes in high-stakes testing within the context of progress testing.

Methods

Ethical approval

The Ethics Committee of the University of Cologne raised no concerns regarding this study or the publication of the results. Only anonymized data were processed and analyzed.

Procedure and Participants

The records of a total of $N_{\text{graduates}}=1240$ medical students who took their national licensing examination between fall 2010 and summer 2014 in Cologne were included in the analysis. The records from the national licensing examination were then matched to the progress test database. Due to regulations in Germany, students do not necessarily take their national licensing examination at the medical school they studied. Thus, the number of students for whom data was available for both the national licensing examinations and from progress tests was $N_{\text{fulldata}}=990$.

Educational Context

In Germany, it takes six years and three months of medical education to practice medicine. To earn a medical degree, three summative state examinations have to be successfully completed. At the University of Cologne, the curriculum is divided into three parts. Academic years one and two are the preclinical phase (i.e., semester 1 to 4), the second part is clinical phase (academy year three to five, i.e., semesters 5 to 10) and year six is the practical year. The first state examination can be successfully completed after the preclinical phase at the earliest (i.e., after the fourth semester). It includes a written and an oral part. After the 3-year clinical phase comes the second state examination. In the written examinations all clinical disciplines of medicine are reviewed. After passing the second examination, the practical year begins and closes with the third state examination.

At the University of Cologne, the Berlin Progress Test (BPT) is administered once per semester from academy year one to five. In total, students take up to ten progress tests across within their medical training. Participation has been mandatory for all medical students since 2003. However, students studying abroad or doing their internships in another city than Cologne are exempt from this rule. The BPT is used as a feedback tool and is designed to measure the gradual increase of knowledge for students from semester to semester and to show them possible subject-specific deficits. In Cologne, the BPT is conducted as computer-based assessment. Detailed feedback is provided immediately after students complete the test. Since the BPT is administered at more than 15 facilities in Germany and Austria, feedback to students regarding their relative performance in relation to both their cohort and to the students from all participating institutions is possible. Each student receives an individual evaluation at the end of the semester, containing this information.

Measures

Progress Test

The BPT is developed at Charité – Universitätsmedizin Berlin and is designed as a formative assessment covering a broad range of medical subjects. Every semester about 11,000 students from German-speaking countries take part in the BPT. The BPT consists of 200 interdisciplinary multiple-choice (MC) questions in single-best-answer format. Specifics on the blueprint and subject areas covered can be found in Nouns & Georg (2010). Items principally align with the subject areas addressed in the national licensing exams. The test score is the number of correctly answered items minus the number of incorrectly answered items. If they are unsure about the correct answer to a question, students are supposed to choose the “don’t-know”-option which doesn’t lead to a deduction of points.

National licensing examination results

The state-administered national licensing examinations in Germany take place after the second and the fifth academic year with an additional third examination after the internships in the sixth year. Performances on these national licensing examinations are given as grades with ‘1’ signifying the best performance and ‘4’ as the lowest pass grade with decimals steps between these grades. The first step of the national licensing focuses on knowledge in the pre-clinical domain. This part is divided into two components: a written and an oral examination. The written part consists of four separate exams of 2 hours each and covers anatomy, biochemistry, physiology and medical psychology and sociology. Questions are in multiple-choice, short-answer and short-essay answering format. The oral part is intended to cover understanding of biomedical principles. The two parts of the first national licensing examination are then combined into a single grade. The second step of the national licensing is a multiple-choice examination after the fifth academic year, covering all clinical subjects (i.e., internal medicine, surgery, neurology, etc.) and typically consisting of 320 items. The third national licensing examination is an oral examination with a strong practical focus and takes place in a clinical setting. We use the results from these three national licensing examinations as three distinct variables in the current study. For a more intuitive understanding of the correlations in the structural equation models, we re-scaled these three variables so that higher values signify better performances.

Data Processing and Statistical Analyses

The R Language for Statistical Computing (R Core Team 2016) is used for all data processing and structural equation modelling (Rosseel 2012). Missing data is handled using Full Information Maximum Likelihood (FIML) estimation. In this context, we argue that the assumption of data missing at random (MAR) holds. The interested reader may refer to Graham (2009), Rubin (1976), or Schafer & Graham (2002) for further details on classifications and handling of missing data.

In order to investigate our research question, we fit a latent growth model (LGM) within the framework of structural equation modelling. The crucial feature of latent growth curve models is that they enable the researcher to disentangle levels of proficiency (intercepts) from the steepness of growth (slopes). For determining model fit, we use the recommendations by Hu & Bentler (1999) and Kline (2011) and employ the standardized root mean squared residual (SRMR < .08), the Tucker-Lewis Index (TLI > .95), the Comparative Fit Index (CFI > .95), and the root mean squared error of approximation (RMSEA < .06).

Our analytic procedure employs a two-step approach. First, we estimate separate latent growth models in order to investigate how the developmental pattern across the course of study can be described most adequately. This is because it is not readily evident how far the pre-clinical (semesters 1-4) and the clinical phases (semesters

5 to 10) reflect a continuous development of knowledge or whether those phases should be conceived of as qualitatively different stages in medical training (Schmidt & Rikers 2007). Therefore, we first fit three distinct growth models: First, a model assuming a continuum of growth from the first to the last academic year. Second, a model in which separate, but correlated, growth processes for the pre-clinical phase (semesters 1-4) and the clinical phase (semesters 5-10) are specified. Third, we fit a variant of the previous model which allowed for different patterns of growth from semester 1 to semester 5 on the one hand and from semester 6 to semester 10 on the other hand. Since semester 5 marks the transition from the pre-clinical to the clinical phase, and progress test were administered in the first week of the semester, one might argue that this transition phase is more similar to the pre-clinical part than to the clinical part. In the first step, we investigate how these phases can be modeled most adequately and fit the three models to the data. In the second step, we combine the model deemed most adequate and a latent variable model that represents performance on the national licensing examinations. Thus we are able to model the relation between students' patterns of development and their performance in high-stakes national licensing exams.

Results

Descriptive statistics

Table 1 presents means, standard deviations, and bivariate correlations for all variables included in the analyses. Data were used from N=990 individual students. Pearson correlations are calculated using pairwise complete cases. Correlations between successive measurement occasions on the progress test range between a maximum of $r_{(9,10)}=.75$ and a minimum of $r_{(1,10)}=.23$. For the interrelation for only the national licensing exam, the highest correlation is between the basic sciences examination and the written clinical examination ($r=.56$) and the lowest is a correlation of $r=.38$ between the basic sciences examination and the oral clinical examination.

Latent Growth Model

How can pre-clinical and clinical growth trajectories be modeled most adequately?

The models for describing the growth trajectories in semesters 1 to 10 are presented in Table 2. The model with the best overall fit is Model 2, which indicates distinct - but interrelated - patterns of growth in the pre-clinical and clinical phase.

Performance on national licensing exams as a predictor of growth

The latent variable for the national licensing examinations, using performances in the preclinical and two clinical licensing examinations as indicators, had, by definition (Bentler 1990), a perfect model fit (CFI=1, TLI=1, RMSEA=0.00, SRMR=0.00). Subsequently, we estimated a full model that combined the latent growth models for the pre-clinical and clinical phases with the latent variable for the national licensing examination. The results from this joint modelling are given in Figure 1. The latent growth model indicates that relations to performance in national licensing exams were moderate to strong for both intercept and slope factors in the clinical phase ($\rho_{(Ic,NLX)}=.66$; $\rho_{(Sc,NLX)}=.62$), and low to moderate for the slope and intercept in the pre-clinical phase ($\rho_{(Ip,NLX)}=.25$; $\rho_{(Sp,NLX)}=.43$). We observed a moderate to strong relation between intercepts in the pre-clinical phase and the intercept in the clinical phase of medical training ($\rho_{(Ip,Ic)}=.58$), at the same time, slopes within the pre-clinical phase were substantially correlated to the intercepts in the clinical phase ($\rho_{(Sp,Ic)}=.63$). Model fit for this model was deemed adequate (CFI=0.97; TLI=0.97; RMSEA=0.44; RMSEA_(90%CI)=0.037, 0.050; SRMR=0.040).

For checking the robustness of the overall model results, we fitted several variants of this model to investigate to what extent parameter estimates would be affected by changes to the specifications and the data. First, we fixed rather low correlations (i.e., ranging between -.2 and .2) to zero. In this case, the magnitude of the freely estimated coefficients are largely unchanged. Second, we estimated two separate growth models for either the pre-clinical or clinical phase in conjunction with the latent variable for performance on the national licensing exams. Again, the separate analyses result in highly similar estimates: that is, for the clinical phase, correlations of $r_{(CLINint, NLX)}=0.59$ and $r_{(CLINslope, NLX)}=0.65$ for the relation between the national licensing examination latent variable and the intercept and slope factors indicating retention of knowledge, respectively. This relation is weaker for the pre-clinical phase, with correlations of $r_{(PRECLINint, NLX)}=0.16$ and $r_{(PRECLINslope, NLX)}=0.53$. The observed deviation in parameter estimates from the full model and the separate models are negligible with differences in correlations of about $r_{(DIFF)}=.02$. Third, we constrained the sample size, using only students who sat the test a minimum amount of time (more than 40 minutes or more than one hour). Again, parameter estimates were hardly affected by using these differently restricted datasets and fit indices supported the adequacy of the models. Finally, we altered the specification of the national-licensing examination latent variable. In this variant, only the two indicators from the final examination were included. In this case, the correlation between the slope factor and the covariance of the performances on the two national licensing examinations rose to $r_{(CLINslope, NLX)} = 0.72$.

Discussion

In this study we focus on two critical components of the development of knowledge as measured by progress tests and their relation to students' performance in a high-stakes national licensing examination. While several studies address progress tests' general ability to reflect increasing levels of performance over time, these studies have only provided limited evidence for the ability of progress tests to actually reflect between-person differences in the rate of growth of medical knowledge over time. We investigate to what extent between-person variation in levels of performance and the rates of gains in performances can be regarded as distinct factors in describing students learning trajectories. We furthermore analyze how to which extent these components are related to an external criterion, in this case, performance on high-stakes national licensing examinations.

Our results provided evidence for progress tests' suitability to monitor students' retention of knowledge across the course of medical training and highlight that both intercept and slope factors contribute in explaining later performances on high-stakes tests. Finally, the results provide an account for the more general development of students' medical knowledge in medical training. Our analyses suggest that both, students level of knowledge when entering medical school and the steepness of growth are related to levels of knowledge when entering the clinical phase of medical training. Then, initial levels and gains of knowledge characterize students' development of medical knowledge and are substantially related to later performances on national licensing examinations. This general developmental pattern is well in-line with current accounts of the development of expertise in medicine (Kulasegaram et al. 2013).

The results may also offer a new perspective on the use of results from progress tests for benchmarking efforts. For instance, the effectiveness of an instructional approach might be captured by the ability to "lift" comparably low-performing students to the level of students with higher initial ability. Thus, the pace of growth of knowledge attributable to the curricular environment may be an additional useful item to use in inter-institutional comparisons. At least, analyses of the development of students within a particular cohort could be used as a tool for early diagnosis of performance issues in that particular group of students and an impetus to take remediation measures as soon as possible. In contexts where financial support of medical schools is, at least partly, linked to performance indicators, information obtained from progress tests may indeed constitute an additional criterion for judging the effectiveness of a particular institution or curriculum. However, such an approach may change the purpose of the assessment and could have adverse effects--an issue, for instance, evident in the discussion on value-added modelling (i.e., analyses of the additional effect of a particular school

or an individual teacher on students' learning trajectories) in the context of measuring school efficiency (Raudenbush 2004; Anderman et al. 2010).

This study has several limitations. An immediate issue to be noted is that the study presented here is situated in a specific educational context where progress tests are used only as a formative assessment, that is, as a feedback tool for both students and curriculum planners. It is, to this point, unclear how results obtained here can be generalized to other contexts that use progress tests as high-stakes summative assessments. Finally, how far it is appropriate to combine the three distinct parts of the national licensing examination into one latent variable might be a matter of discussion. While the inter-correlations and model-fitting procedures justify this approach on a technical level, the exams indeed are several years apart from each other and cover contents from different subject areas.

Despite these limitations, this study goes beyond the existing literature on progress testing and provides an additional argument for the validity this approach. Most studies investigate growth across the course of study only rather imprecisely, that is, averaging across students within academic years. This approach might also hide a highly interesting component of progress tests: that is, their ability to monitor students' growth of knowledge across medical training on the level of the individual. In this respect, our study adds a crucial point to the literature by not only showing that a substantial amount of variation can be attributed to different rates of growth of knowledge across medical training but also that those rates of growth are related to an external criterion, that is, performances on a high stakes licensing examination.

Figure 1 Combined model for separate growth factors in pre-clinical (PT1-4) and clinical phase (PT5-10) of training and their relation to performances on national licensing examinations (NLX1-3). Two-headed arrows represent correlations, single-headed arrows represent loadings. As usual in latent growth curve models, all loadings on the intercept factors are fixed to 1 and the loadings on the slope factor are fixed from 0 to 3 for the preclinical phase, and 0 to 5 for the clinical phase. Error variances for the manifest variables PT1-4 and PT5-10 are fixed to be equal. The first loading for the latent variable for performances on national licensing examinations is fixed at 1. The other two loadings are estimated freely and intercepts and variances are estimated for NLX indicators.

302
303
304

Table 1: Descriptive statistics

			Pearson correlations (pairwise complete cases)											
	M	SD	PT 1	PT 2	PT 3	PT 4	PT 5	PT 6	PT 7	PT 8	PT 9	PT 10	NLX 1	NLX 2
PT 1	4.26	9.60	-											
PT 2	7.48	10.23	0.65	-										
PT 3	11.38	9.95	0.55	0.55	-									
PT 4	16.22	11.47	0.39	0.50	0.59	-								
PT 5	23.47	18.90	0.31	0.44	0.48	0.60	-							
PT 6	29.02	17.58	0.40	0.48	0.56	0.64	0.65	-						
PT 7	36.42	21.38	0.29	0.37	0.49	0.56	0.66	0.69	-					
PT 8	44.55	21.77	0.29	0.36	0.51	0.56	0.67	0.68	0.69	-				
PT 9	52.17	23.96	0.27	0.32	0.43	0.52	0.61	0.65	0.66	0.72	-			
PT 10	58.25	27.87	0.23	0.29	0.44	0.53	0.62	0.68	0.68	0.69	0.75	-		
NLX 1	2.55	0.68	-0.10	-0.10	-0.26	-0.37	-0.35	-0.39	-0.43	-0.46	-0.47	-0.49	-	
NLX 2	2.72	0.74	-0.15	-0.19	-0.26	-0.33	-0.38	-0.42	-0.47	-0.50	-0.50	-0.55	0.56	-
NLX 3	1.82	0.78	-0.08	-0.12	-0.18	-0.22	-0.29	-0.32	-0.34	-0.37	-0.38	-0.40	0.38	0.44

Note M, mean; SD, standard deviation; PT1-10 signify the progress tests taken in the first to the tenth semester. NLX 1 is the national licensing exam in the pre-clinical phase, NLX 2 the written national licensing exam in the clinical phase and NLX 3 final oral examination.

Table 2: Models for learning trajectories across medical training

	CFI	TLI	RMSEA (90% CI)	SRMR	AIC	BIC
Model 1: One continuum (Semester 1-10)	0.85	0.86	0.11 (0.10 – 0.12)	0.099	54857.872	54931.322
Model 2: Separate phases (Semester 1-4 and 5-10)	0.99	0.99	0.035 (0.025 – 0.045)	0.027	54329.487	54447.008
Model 3: Separate phases (Semester 1-5 and 6-10)	0.95	0.94	0.072 (0.063 – 0.080)	0.064	54487.509	54605.029

Note CFI, Comparative Fit Index; TLI, Tucker-Lewis Index; SRMR, standardized root mean squared residual; RMSEA, root mean squared error of approximation; AIC, Akaike Information Criterion; BIC, Bayes Information Criterion

References

- Aarts R, Steidel K, Manuel BAF, Driessen EW. 2010. Progress testing in resource-poor countries: a case from Mozambique. *Med Teach*. 32:461–463. eng. doi:10.3109/0142159X.2010.486059.
- Ali K, Coombes L, Kay E, Tredwin C, Jones G, Ricketts C, Bennett J. 2016. Progress testing in undergraduate dental education: the Peninsula experience and future opportunities. *Eur J Dent Educ*. 20:129–134. eng. doi:10.1111/eje.12149.
- Anderman EM, Anderman LH, Yough MS, Gimbert T BG. 2010. Value-Added Models of Assessment: Implications for Motivation and Accountability. *Educational Psychologist*. 45:123–137. doi:10.1080/00461521003703045.
- Bentler PM. 1990. Comparative fit indexes in structural models. *Psychological Bulletin*. 107:238–246. doi:10.1037/0033-2909.107.2.238.
- Black P, Wiliam D. 2009. Developing the theory of formative assessment. *Educ Asse Eval Acc*. 21:5–31. doi:10.1007/s11092-008-9068-5.
- Blake, J. M., Norman, G. R., Keane, D. R., & Mueller, C. B. 1996. Introducing progress testing in McMaster University's problem-based medical curriculum: Psychometric properties and effect on learning. *Academic Medicine*, 71(9), 1002-1007.
<http://dx.doi.org/10.1097/00001888-199609000-00016>
- Brauer DG, Ferguson KJ. 2015. The integrated curriculum in medical education: AMEE Guide No. 96. *Med Teach*. 37:312–322. eng. doi:10.3109/0142159X.2014.970998.
- Coombes L, Ricketts C, Freeman A, Stratford J. 2010. Beyond assessment: feedback for individuals and institutions based on the progress test. *Med Teach*. 32:486–490. eng. doi:10.3109/0142159X.2010.485652.
- Dijksterhuis MGK, Jozwiak I, Braat DDM, Scheele F. 2013. An exploratory study into the effect of time-restricted internet access on face-validity, construct validity and reliability of postgraduate knowledge progress testing. *BMC Med Educ*. 13:147. eng. doi:10.1186/1472-6920-13-147.
- Dijksterhuis MGK, Scheele F, Schuwirth LWT, Essed GGM, Nijhuis JG, Braat DDM. 2009. Progress testing in postgraduate medical education. *Med Teach*. 31:e464-8. eng. doi:10.3109/01421590902849545.
- Freeman A, van der Vleuten C, Nouns Z, Ricketts C. 2010. Progress testing internationally. *Med Teach*. 32:451–455. eng. doi:10.3109/0142159X.2010.485231.
- Graham, J. W. 2009. Missing data analysis: Making it work in the real world. *Annual review of psychology*, 60, 549-576

- Hu, L. T., & Bentler, P. M. 1999. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal*, 6(1), 1-55.
- Joiner AB, Abbott S, Longson D. 2016. Is Performance-Based Progress Testing in Psychiatry Feasible? *Acad Psychiatry*. 40:790–795. eng. doi:10.1007/s40596-015-0464-0.
- Kline, R. 2011. *Principles and Practice of Structural Equation Modeling*. New York: The Guilford Press
- Kulasegaram, K. M., Grierson, L. E., & Norman, G. R. (2013). The roles of deliberate practice and innate ability in developing expertise: evidence and implications. *Med Educ*, 47(10), 979-989.
- Leppink J. 2015. Data analysis in medical education research: a multilevel perspective. *Perspect Med Educ*. 4:14–24. eng. doi:10.1007/s40037-015-0160-5.
- Muijtjens AMM, Timmermans I, Donkers J, Peperkamp R, Medema H, Cohen-Schotanus J, Thoben A, Wenink ACG, van der Vleuten CPM. 2010. Flexible electronic feedback using the virtues of progress testing. *Med Teach*. 32:491–495. eng. doi:10.3109/0142159X.2010.486058.
- Neeley SM, Ulman CA, Sydelko BS, Borges NJ. 2016. The Value of Progress Testing in Undergraduate Medical Education: A Systematic Review of the Literature. *Med.Sci.Educ*. 26:617–622. doi:10.1007/s40670-016-0313-0.
- Nouns Z, Schaubert S, Witt C, Kingreen H, Schüttpeitz-Brauns K. 2012. Development of knowledge in basic sciences: a comparison of two medical curricula. *Med Educ*. 46:1206–1214. eng. doi:10.1111/medu.12047.
- Nouns ZM, Georg W. 2010. Progress testing in German speaking countries. *Med Teach*. 32:467–470. eng. doi:10.3109/0142159X.2010.485656.
- Preacher KJ. 2008. *Latent growth curve modeling*. Los Angeles, Calif., London: SAGE (Sage university papers series. Quantitative applications in the social sciences; 07-157). ISBN: 978-1412939553.
- R Core Team. 2016. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Available from: <https://www.r-project.org/>.
- Raudenbush SW. 2004. What Are Value-Added Models Estimating and What Does This Imply for Statistical Practice? *Journal of Educational and Behavioral Statistics*. 29:121–129.
- Rosseel Y. 2012. lavaan: An R Package for Structural Equation Modeling. *J. Stat. Soft*. 48. doi:10.18637/jss.v048.i02.
- Rubin, D. B. 1976. Inference and missing data. *Biometrika*, 63(3), 581-592.
- Schafer, J. L., & Graham, J. W. 2002. Missing data: our view of the state of the art. *Psychological methods*, 7(2), 147.

372 Schmidt HG, Rikers RMJP. 2007. How expertise develops in medicine: knowledge encapsulation and illness
 373 script formation. *Med Educ.* 41:1133–1139. eng. doi:10.1111/j.1365-2923.2007.02915.x.

374 Schuwirth LWT, van der Vleuten CPM. 2011. General overview of the theories used in assessment: AMEE
 375 Guide No. 57. *Med Teach.* 33:783–797. eng. doi:10.3109/0142159X.2011.611022.

376 Tio RA, Schutte B, Meiboom AA, Greidanus J, Dubois EA, Bremers AJA. 2016. The progress test of medicine:
 377 the Dutch experience. *Perspect Med Educ.* 5:51–55. eng. doi:10.1007/s40037-015-0237-1.

378 van der Vleuten CPM, Heeneman S. 2016. On the issue of costs in programmatic assessment. *Perspect Med*
 379 *Educ.* 5:303–307. eng. doi:10.1007/s40037-016-0295-z.

380 Wiliam D, Black P. 1996. Meanings and Consequences: A Basis for Distinguishing Formative and Summative
 381 Functions of Assessment? *British Educational Research Journal.* 22:537–548.

382 Wimmers PF, Lee M. 2015. Identifying longitudinal growth trajectories of learning domains in problem-based
 383 learning: a latent growth curve modeling approach using SEM. *Adv Health Sci Educ Theory Pract.* 20:467–
 384 478. eng. doi:10.1007/s10459-014-9541-5.

385 Wrigley W, van der Vleuten CPM, Freeman A, Muijtjens A. 2012. A systemic framework for the progress test:
 386 strengths, constraints and issues: AMEE Guide No. 71. *Med Teach.* 34:683–697. eng.
 387 doi:10.3109/0142159X.2012.704437.