Aus dem Institut für Neuropsychologie und Klinische Psychologie
des Zentralinstituts für Seelische Gesundheit Mannheim

Wissenschaftliche Direktorin: Prof. Dr. Dr. h.c. Dr. h.c. Herta Flor

The Utility of Low-Stakes Assessment with the Example of the Berlin Progress Test

Kumulative Habilitationsschrift

zur Erlangung der Venia Legendi für das Fach
Medizindidaktik

der Hohen Medizinischen Fakultät Mannheim
der Ruprecht-Karls-Universität Heidelberg

vorgelegt von
Dr. Katrin Schüttpelz-Brauns
aus Mannheim
2020

Contents

Chapter 1

Introduction

Low-stakes assessment is still an unfamiliar concept to most people, but it has received increasing attention in recent years.

In contrast to high-stakes assessments where grades are given to test-takers, *"there are typically no consequences associated with student performance"* in low-stakes assessment (Wise & DeMars, 2005, p. 2). In a culture where grades are the aim of a course or a study year there seems to be no rationale for low-stakes assessments.

However, this kind of assessment serves several purposes: for students as formative assessment, for faculty as evaluation tool, for policy as large-scale assessment and for society as research tool.

As *formative assessment*, they provide feedback to students and teachers that guides the learning process (Dunn & Mulvenon, 2009; Martinez & Lipson, 1989) and can increase learning effects (Black & William, 1998; Hattie & Timperley, 2007; Kluger & DeNisi, 1996; Shute, 2008). Formative assessments are an important part of self-regulated learning (Ecclestone, 2010; Irons, 2008; White & Gruppen, 2010; Wood, 2010) and enable deep learning (Nicol & Macfarlane-Dick, 2006; Rushton, 2005). Studying formative (low-stakes) assessment can help to understand how self-regulated learning functions.

As *evaluation tool*, data from low-stakes assessments provide information on classroom teaching, as well as for academic discourse (Dunn & Mulvenon, 2009). As evaluation tool in educational institutions, low-stakes assessments objectively document students' achievements (Cole & Osterlind, 2008). Examples for using standardized tests to evaluate

students' academic success are the College BASE, the Collegiate Assessment of Academic Proficiency (CAAP), the Measure of Academic Proficiency and Progress (MAPP), and the Collegiate Learning Assessment (CLA) (Cole, 2007). Student evaluations can have direct consequences on teachers (Rutkowski & Wild, 2015).

Large-scale assessments, like the Programme for International Student Assessment (PISA), Trends in International Mathematics and Science Study (TIMSS), National Assessment of Educational Progress (NAEP) or the Pan-Canadian Assessment Program (PCAP), are sometimes low-stakes for participants (Brunner, Artelt, Krauss, & Baumert, 2007; Copp, 2018). They provide information for benchmarking and are part of the quality management of educational programs (Campbell, Voelkl, & Donahue, 1998; Copp, 2018; OECD, 1999, 2003). Large-scale assessments inform policy. As so-called evidence-based education policy the results are used e.g. to initiate educational reforms (Breakspear, 2012; Butler & Adams, 2007; Copp, 2018; Fullan, 2009).

Furthermore, low-stakes assessments are used in research. As *research tool*, they help to provide findings for the research community, faculties and society. In this context low-stakes assessment is used, for example, to gain insights on knowledge acquisition, like long-term retention (Bae, Therriault, & Redifer, 2018; Bell et al., 2008; Custers & ten Cate, 2011; Larsen, Butler, & Roediger III, 2013; Roediger III & Butler, 2011, etc.), prior knowledge activation (Alvermann, Smith, & Readence, 1985; Crooks & Alibali, 2013; Kostons & van der Werf, 2015; Wetzels, Kester, van Merriënboer, & Broers, 2011, etc.) or delayed retention learning (Chang, 2017; Haynie III, 1994; Ramraje & Sable, 2011, etc.). A large body of research using low-stakes assessments is conducted on educational interventions, like single teaching methods (Dankbaar et al., 2017; Löffler et al., 2011; Raupach et al., 2015; Renkl, Stark, Gruber, & Mandl, 1998; Schmidmaier et al., 2011; Seybert & Barton, 2007, etc.) or training programs (Finch, 1999; Lambert, 2001; Löwe et al., 2008; Prince et al., 2003; Raman et al., 2010, etc.).

Low-stakes assessments suffer from high variation in test-taking effort (Hosch, 2012; Setzer, Wise, van den Heuvel, & Ling, 2013; Wise & DeMars, 2005; Wise, 2009), which means *"giving one's best effort to the test"* (Wise & DeMars, 2005, p. 2). Test scores,

therefore, do not only reflect ability (Barry, Horst, Finney, Brown, & Kopp, 2010; Eklöf & Knekta, 2017; O'Neil, Sugrue, & Baker, 1995; Sundre & Kitsantas, 2004; Wise & Kong, 2005; Wolf & Smith, 1995). Participants with higher levels of test-taking effort outperform participants with lower test-taking effort (Baumert & Demmrich, 2001; Cole, Bergin, & Whittaker, 2008; Liu, Bridgeman, & Adler, 2012; Thelk, Sundre, Horst, & Finney, 2009; Wise & DeMars, 2005). Although they seem low-stakes for participants, low-stakes assessments can have severe consequences for teachers, faculty, institutions or policy (Breakspear, 2012; Cole, 2007; Cole & Osterlind, 2008). If test-taking effort is not taken into account, the validity of results can be threatened (Akyol, Krishna, & Wang, 2018; Brown & Walberg, 1993; Butler & Adams, 2007; Eklöf, 2010; Penk, 2017; Thelk et al., 2009; Wise & DeMars, 2005; Wolf & Smith, 1995). As a consequence, low-stakes assessments may not serve their purposes properly, however, low-stakes assessment is only useful if it serves its purpose.

About 15 years ago I started working on a low-stakes test, the Berlin Progress Test (BPT), and some time later I conducted research on it. Progress tests can serve all of the purposes described above and they represent both moderate-stakes and low-stakes assessment. Therefore research findings concerning moderate-stakes versus low-stakes progress tests can be compared and thus the special aspects of low-stakes assessments can be worked out.

In this work I first introduce progress testing: why progress tests were developed (s. 2.1.1), what they look like (s. 2.1.2), where they are used (s. 2.1.3), how the purposes of low-stakes assessments are fulfilled (s. 2.1.4) and what the stakes of progress tests mean (s. 2.1.5). In section 2.2, I discuss moderate- and low-stakes progress tests in terms of the findings for each component of the model of Utility of Assessment Methods (van der Vleuten, 1996) to outline the special features of low-stakes assessment, which are the subject of the studies in my research.

Low-stakes assessments are a big field with many facets. I therefore want to limit the subject of the research to progress tests in medical education.

Chapter 2

Background

2.1 Progress Testing

2.1.1 Origin of Progress Testing

Progress tests were independently invented in the 1970s at the University of Missouri in the United States (Arnold & Willoughby, 1990) and at the University of Limburg in the Netherlands (van Berkel, 1990). Both medical schools had implemented problem-based curricula in their undergraduate medical training meant to encourage deeper learning styles, but the influence of the assessments at the end of the teaching blocks prevented deeper learning strategies and encouraged rote memorization (Blake et al., 1996; van der Vleuten, Verwijnen, & Wijnen, 1996). This was the reason why progress tests were developed.

To break the link between assessment and curriculum (Albanese & Case, 2016; Blake et al., 1996), progress tests are tailored to the end objectives of the curriculum (Freeman, van der Vleuten, Nouns, & Ricketts, 2010; Nouns & Brauns, 2008; van der Vleuten et al., 1996) and sample questions are taken from the whole field of medicine which is taught in undergraduate medical education (Albanese & Case, 2016; Blake, Norman, & Smith, 1995; Blake et al., 1996; Tio et al., 2016; Wrigley, van der Vleuten, Freeman, & Muijtjens, 2012) and expected of students at graduation (Arnold & Willoughby, 1990; Nouns & Brauns, 2008). This makes preparation for progress tests difficult (Albanese & Case,

2016; Freeman et al., 2010), especially with memorization techniques (Albanese & Case, 2016; Pugh & Regehr, 2016; van Berkel, Nuy, & Geerligs, 1995). Since students cannot prepare for the test, it spontaneously captures long-term knowledge (Heeneman, Schut, Donkers, van der Vleuten, & Muijtjens, 2017; Nouns & Brauns, 2008; Schuwirth & van der Vleuten, 2012; van Berkel et al., 1995) and has advantages for students who use deeper learning styles (Albanese & Case, 2016; Freeman et al., 2010; Pugh & Regehr, 2016; van Berkel et al., 1995).

2.1.2 Description of Progress Tests

Progress tests are administered repeatedly throughout the whole curriculum, regardless of the semester level (Albanese & Case, 2016; Nouns & Brauns, 2008; van der Vleuten et al., 1996; Wrigley et al., 2012). All students sit the same test. The number of tests administered varies between one and four per study year (Freeman et al., 2010). A new test is compiled from a question bank for each test administration (Albanese & Case, 2016) and represents the whole of the curriculum (Blake et al., 1995) according to a blueprint that specifies the number of questions among different content areas, such as domains and subjects (Albanese & Case, 2016; Coombes, Ricketts, Freeman, & Stratford, 2010; Nouns & Brauns, 2008; Tio et al., 2016). However, not all medical schools use a blueprint (Findyartini et al., 2015; Tomic, Martins, Lotufo, & Benseñor, 2005). Over time the question banks have grown large. As an example, the question bank for the Maastricht progress test consisted of 15,000 items in 1996 (van der Vleuten et al., 1996).

Almost all progress tests consist of multiple-choice questions with one best answer. There are also progress tests with true/false questions (Albanese & Case, 2016; van der Vleuten et al., 1996) and case scenarios with open responses (Albanese & Case, 2016; Rademakers, Ten Cate, & Bär, 2005). The number of questions per progress test ranges from 120 (Findyartini et al., 2015) to 250 (van der Vleuten et al., 1996). The questions contain patient vignettes or complex medical problems to apply medical knowledge even when asking about basic medical sciences (Albanese & Case, 2016; Ricketts, Freeman, Pagliuca, Coombes, & Archer, 2010).

Questions are written by faculty and go through a quality cycle of writing, reviewing and revising (Nouns & Georg, 2010; Osterberg, Kölbel, & Brauns, 2006; Tio et al., 2016; van der Vleuten et al., 1996). At some schools question statistics are sent to authors after the test for question revision according to this feedback (Nouns & Georg, 2010; Osterberg et al., 2006; van der Vleuten et al., 1996). A large number of faculty is involved in question writing and reviewing, as well as in the final compilation of the progress test (Albanese & Case, 2016). Because of the continuing quality cycle of question writing, reviewing, revising, using them in the test, revising them again after test administration, almost all questions are changed before being administered in a new test (Albanese & Case, 2016).

Because students are not able to answer all end-of-the-curriculum questions, especially at the beginning of undergraduate training, a "don't know" option is included as a possible answer (Nouns & Brauns, 2008; Tio et al., 2016; van der Vleuten et al., 1996). Not all medical schools use this option (Tomic et al., 2005). The use of the "don't know" option is discussed in more detail in Wrigley et al. (2012). Additionally, most schools use formula scoring, which means that the test score is calculated as number of correct answers minus number of incorrect answers. This is meant to prevent students from guessing (Albanese & Case, 2016; van der Vleuten et al., 1996).

Progress tests provide detailed feedback not only to students, but also to groups and institutions, like mentors, teachers in the program, departments and curriculum committees, and show performance of students, cohorts, curricula and institutions (Albanese & Case, 2016; Coombes et al., 2010; Tio et al., 2016; van der Vleuten et al., 1996; Wrigley et al., 2012).

There are some review articles that describe progress tests in more detail (Albanese & Case, 2016; Neeley, Ulman, Sydelko, & Borges, 2016; Wrigley et al., 2012; Plessas, 2015).

2.1.3 Spreading of Progress Tests

Progress tests in medical undergraduate training are used all over the world: in Africa it is Mozambique (Aarts, Steidel, Manuel, & Driessen, 2010) and South Africa (Freeman et al., 2010), in North America it is Canada (Blake et al., 1995) and the United States

(Arnold & Willoughby, 1990), in South America it is Brazil (da Rosa et al., 2017; Tomic et al., 2005), in Asia it is Indonesia (Findyartini et al., 2015; Mardiastuti & Werhani, 2011) and Saudi Arabia (Al Alwan et al., 2011; Soliman, Al-Shaikh, & Alnassar, 2016), in Oceania it is New Zealand (Lillis et al., 2014), and in Europe it is Austria (Nouns & Georg, 2010), Finland (Freeman et al., 2010), Germany (Nouns & Georg, 2010), Ireland (Given, Hannigan, & McGrath, 2016), the Netherlands (Rademakers et al., 2005; Tio et al., 2016; van Berkel et al., 1995; van der Vleuten et al., 1996) and the United Kingdom (Freeman & Ricketts, 2010).

Most progress tests are implemented in undergraduate medical training (Lillis et al., 2014) but also in disciplines like anatomy (Hanß, 2013), in undergraduate dentistry (Ali et al., 2016; Bennett, Freeman, Coombes, Kay, & Ricketts, 2010; Freeman et al., 2010; Kirnbauer et al., 2018; Wrigley et al., 2012), veterinary medicine (Siegling-Vlitakis et al., 2014), and psychology (Schaap, Schmidt, & Verkoeijen, 2012; Wrigley et al., 2012), as well as postgraduate training in internal medicine (Pugh, Touchie, Wood, & Humphrey-Murto, 2014), obstetrics & gynecology (Dijksterhuis et al., 2009), osteopathic medicine (Portanova et al., 2000), and radiology (Ravesloot et al., 2012; Rutgers et al., 2018).

This list refers only to the published information. There may well be progress tests at other locations and in other disciplines, and in undergraduate as well as postgraduate programs.

2.1.4 Application of Progress Tests

Because progress tests are administered regularly, it is said that they can enhance learning (Custers, 2008; Larsen, Butler, & Roediger III, 2008; van der Vleuten, Freeman, & Collares, 2018). Some studies from the field of psychology have already shown the effect of repeated testing on learning (Bangert-Drowns, Kulik, & Kulik, 1991; Karpicke & Roediger III, 2007; McDaniel, Anderson, Derbish, & Morrisette, 2007; Roediger III & Butler, 2011).

When a comprehensive domain of knowledge is tested repeatedly, it cannot be studied the night before, thereby promoting long-term knowledge and knowledge retention (van

der Vleuten et al., 1996).

As performance is accumulated over several test occasions, students feel less stressed or anxious when sitting the test (Albanese & Case, 2016; Blake et al., 1996; Reynolds & Kostich, 2017; Schuwirth & van der Vleuten, 2012; van der Vleuten et al., 1996).

Progress tests can fulfill all purposes of low-stakes assessments as described in chapter 1.

Progress Tests as Formative Assessments

Detailed feedback on the results is provided to students and mentors cross-sectionally as profile scores and longitudinally as growth curves (Blake et al., 1995, 1996; Coombes et al., 2010; McHarg et al., 2005; Muijtjens et al., 2010; Neeley et al., 2016; Nouns & Brauns, 2008; Ricketts et al., 2010; van der Vleuten et al., 1996). The format can be numerical and/or graphical (Neeley et al., 2016). Results are summarized as averages of the test score, of organ systems, of disciplines, and/or of clusters of disciplines like basic medical sciences, clinical sciences, or behavioral sciences (Muijtjens et al., 2010; Tio et al., 2016; van der Vleuten et al., 1996). Cohort means allow group comparisons (norm-referencing).

Students can use the progress test results to identify strengths and weaknesses (Aarts et al., 2010; Blake et al., 1995, 1996; Given et al., 2016; Muijtjens et al., 2010) and can adapt their learning (Given et al., 2016; Wade et al., 2012; Yielder et al., 2017). Mentors and faculty advisors at some medical schools receive accumulated progress test results after three progress tests (Blake et al., 1995; van der Vleuten et al., 1996)

Progress Tests as Evaluation Tools

Progress tests are used as evaluation tools within a curriculum regarding knowledge acquisition (Peeraer et al., 2009).

Growth in the percentage correct for individual questions on progress tests helps teachers improve their teaching; the average performance in single disciplines helps departments improve (Coombes et al., 2010; De Champlain et al., 2010; Wrigley et al., 2012).

Faculty staff involved in quality management systems use the average performance in

disciplines and organ systems to check the effectiveness of an existing curriculum regarding knowledge acquisition and, if needed, to modify it and then monitor the effectiveness of the modification (Aarts et al., 2010; Al Alwan et al., 2011; Coombes et al., 2010; Findyartini et al., 2015; Nouns & Brauns, 2008; Schmidmaier et al., 2010; Schuwirth, Bosman, Henning, Rinkel, & Wenink, 2010; Tio et al., 2016; van der Vleuten et al., 1996, 2004).

The same information is used by curriculum developers to monitor knowledge acquisition when switching from a traditional to a reformed curriculum (Neeley et al., 2016; Tio et al., 2016; van der Veken, Valcke, De Maeseneer, Schuwirth, & Derese, 2009) or to check the quality of a new curriculum (Neeley et al., 2016; Finucane, Flannery, Keane, & Norman, 2010; Freeman & Ricketts, 2010; Johnson, Khalil, Peppler, Davey, & Kibble, 2014; Tio et al., 2016).

Progress Tests as Large-Scale Assessments

Since questions on progress tests refer to end-of-curriculum objectives, a change in curriculum has no consequence on the progress test, provided the end-of-curriculum objectives do not change (van der Vleuten et al., 1996). This is why progress tests can measure change in knowledge acquisition regardless of the curriculum. It does not matter how unique a curriculum is or what methods of teaching and learning are used. With progress testing the knowledge acquisition in different curricula can be compared within and across countries (Cecilio-Fernandes, Aalders, Bremers, Tio, & de Vries J., 2018; Muijtjens, Schuwirth, Cohen-Schotanus, Thoben, & van der Vleuten, 2008; Neeley et al., 2016; Schauber & Nouns, 2010; Tio et al., 2016; van der Veken et al., 2009; van der Vleuten et al., 1996, 2004, 2018; Verhoeven et al., 1998, 2005). This provides opportunities for benchmarking, provided all institutions use the same progress test (Muijtjens, Schuwirth, Cohen-Schotanus, & van der Vleuten, 2007; Schuwirth et al., 2010).

Progress Tests as Research Tools

Progress tests are used to do research on knowledge acquisition throughout the curriculum (Boshuizen, van der Vleuten, Schmidt, & Machiels-Bongaerts, 1997; Verhoeven, Verwi-

jnen, Scherpbier, & van der Vleuten, 2002), at different sites (Bianchi, Stobbe, & Eva, 2008), in different disciplines and domains, like basic, behavioral, and/or clinical science (Tomic et al., 2005; van der Vleuten et al., 1996; van Diest et al., 2004), and to analyze the rehearsal effect (Kerfoot et al., 2011).

2.1.5 Stakes of Progress Tests

The stakes of progress tests differ depending on the consequences drawn from the results (Albanese & Case, 2016). Progress tests in Austria, Brazil, Finland, Germany, Indonesia, and Ireland are "purely formative" (Findyartini et al., 2015; Finucane et al., 2010; Freeman et al., 2010; Nouns & Georg, 2010; Tomic et al., 2005). This means results of the performance are given to the students as feedback and guidance for future learning, but the performance itself has no consequences for advancement in the undergraduate training. These progress tests are low-stakes. Some progress tests are used to identify consistently low-performing students. In this case, (accumulated) progress tests lead to enforced remediation or hinder advancement in the undergraduate training (Aarts et al., 2010; Arnold & Willoughby, 1990; Blake et al., 1996; Coelho, Zahra, Ali, & Tredwin, 2019; Lillis et al., 2014; Norman, Neville, Blake, & Mueller, 2010; Tio et al., 2016; van der Vleuten et al., 1996). These progress tests have moderate stakes.

In low-stakes progress tests students do not face negative consequences if they do not perform at their best. Test scores, therefore, do not only reflect ability but also test-taking effort with corresponding effects on test validity (Wise & Kong, 2005; Barry et al., 2010). As a consequence of this high variation, low-stakes progress tests may not work as a feedback instrument, evaluation tool, large-scale assessment or research tool.

Nouns and Georg (2010, p. 468) listed the advantages of a low-stakes progress test and the reason why they don't raise the stakes to prevent the so-called non-serious test-takers: *"Test results are not biased by vast preparation; students are discouraged from collecting test items; tests do not interfere with curriculum; tests are no extra burden for students with a high exam load."* Using low stakes instead of moderate stakes may impact the utility of progress tests. For this reason the utility of low-stakes progress tests has to be

studied.

Still, there is one more aspect that should not be ignored: in educational research self-regulated learning has been one of the major research topics of recent years (Panadero, 2017). In self-regulated learning students use feedback (e.g. from low-stakes assessment) to match actual learning or performance with their learning goals to regulate their learning (Nicol & Macfarlane-Dick, 2006). Progress testing is an example of formative assessment as part of self-regulated learning, because *"The cycle of testing, giving feedback, students using that feedback to direct learning and then retesting is inherent in progress testing."* (Ricketts et al., 2010, p. 515). Studying how low-stakes progress tests function could therefore be a contribution to understanding how self-regulated learning works.

However, we must first take a look at the specifics of low-stakes assessments by comparing moderate- and low-stakes progress tests. This can be done using the Model of Utility of Assessment Methods (van der Vleuten, 1996).

2.2 Utility of Progress Tests

2.2.1 Model of Utility of Assessment Methods

Regardless whether their stakes are high, moderate or low, assessments have to be constructed carefully. Different decisions on the assessment design and implementation strategies have to be made to increase utility and have practical implications. To address this, van der Vleuten (1996) developed a model of "Utility of Assessment Methods". According to this model, utility components, namely reliability, validity, educational impact, acceptability, and cost effectiveness, have to be combined according to the assessment purpose. In the following section, I briefly describe the theoretical components of the utility model followed by the comparison of findings from publications about moderate- and low-stakes progress tests.

2.2.2 Reliability

Theoretical Considerations

An assessment is reliable if scores are reproducible. In classical test theory the observed score is the sum of the true score and some error:

$$X = T + e$$

where X=observed score, T=true score, e=error. Reliability is an indicator of the ratio of true scores in the observed scores and the measurement error (Downing, 2004; Streiner, Norman, & Cariney, 2015). The smaller the measurement error, the higher the reliability. To improve reliability (1) the number of questions in the assessment should be maximized (Downing, 2004) and (2) questions should be written with great care using question writing guidelines (Case & Swanson, 2002; Downing, 2004, 2005; Haladyna, Downing, & Rodriguez, 2010; Ware & Vik, 2009). Poorly crafted items can be prevented by training question authors, providing feasible guidelines and reviewing the questions in a way that includes formal criteria (Albanese & Case, 2016).

Downing (2004) has summarized the most frequent opinions on the sizes of reliable assessment depending on the stakes. It is at least 0.90 for very high-stakes assessments, such as certification examinations in medicine. It falls in the range of 0.80-0.89 for assessments with moderate stakes, such as end-of-course examinations at a medical school and in the range of 0.70-0.79 for assessments with lower or no consequences, such as formative assessments. If progress tests are reliable, (3) their size of reliability should meet these requirements.

Reliability of Progress Tests

Progress tests do have a large number of questions with a minimum of 125 in moderate-stakes progress tests (Freeman & Ricketts, 2010; Wrigley et al., 2012) and a minimum of 120 questions in low-stakes progress tests (Findyartini et al., 2015) (1).

Although training question authors, providing guidelines and reviewing the questions

regarding formal criteria is mentioned in one review of progress testing (Wrigley et al., 2012), it is not described in most of the papers (2). Quality assurance in the item writing process itself is described in moderate-stakes progress tests (Rademakers et al., 2005; Tio et al., 2016) and in low-stakes progress tests (Nouns & Georg, 2010). Providing feedback to the item authors regarding student answers is described in two papers, one describing a moderate-stakes progress test (van der Vleuten et al., 2004) and the other a low-stakes progress test (Nouns & Georg, 2010). This presents an opportunity for training the authors later on.

Studies on moderate-stakes progress tests in undergraduate medical education show that reliability ranges (3)

- for multiple-choice questions, between $\alpha = 0.82$ and $\alpha = 0.94$ for the entire sample (Findyartini et al., 2015; Kerfoot et al., 2011; Swanson et al., 2010; Tio et al., 2016), between $\alpha = 0.46$ and $\alpha = 0.73$ per cohort (Blake et al., 1996; Kerfoot et al., 2011), and between $r_{\text{tt}} = 0.53$ and $r_{\text{tt}} = 0.70$ for test-retest reliability over successive intervals (Blake et al., 1995; Albanese & Case, 2016),

- for true false questions, between $\alpha = 0.78$ and $\alpha = 0.95$ for the entire sample (Aarts et al., 2010; Boshuizen et al., 1997; van der Vleuten et al., 1996), and between $\alpha = 0.66$ and $\alpha = 0.80$ per cohort (Boshuizen et al., 1997; van der Vleuten et al., 1996),

- for open-ended, short-answer questions, between $\alpha = 0.85$ and $\alpha = 0.86$ for the entire sample (Rademakers et al., 2005).

The reliability of low-stakes progress tests ranges between $\alpha = 0.85$ and $\alpha = 0.98$ for the entire sample (Findyartini et al., 2015; Osterberg et al., 2006; Nouns & Georg, 2010) and averages $\alpha = 0.85$ within cohorts of the same level (Nouns & Georg, 2010).

2.2.3 Validity

Theoretical Considerations

An assessment is valid if it measures what it is developed for or, in other words, *"a valid test measures what it is intended to measure"* (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 1999). Assessment validity consists of different types: content validity, construct validity, and criterion validity, subdivided into concurrent and predictive validity.

Content validity examines whether the content is appropriately represented. This can be ensured (1) by using a blueprint that specifies a consistent ratio of questions per organ system and per discipline for each test administration (Plessas, 2015; Wrigley et al., 2012) and (2) if a review committee of content experts discusses the content validity of the assessment questions (Gesellschaft für Medizinische Ausbildung, GMA-Ausschuss Prüfungen and Kompetenzzentrum Prüfungen, Baden-Württemberg & Fischer, 2008; Jünger & Just, 2014; Verhoeven, Verwijnen, Scherpbier, Schuwirth, & van der Vleuten, 1999; Wallach, Crespo, Holtzman, Galbraith, & Swanson, 2006; Ware & Vik, 2009). In progress testing this type of validity is at risk if the questions are not at the required level of difficulty (e.g. undergraduate vs. postgraduate level). Therefore, review committees engaged in progress testing have to ensure that only graduate knowledge is queried in the questions (Albanese & Case, 2016; Plessas, 2015; Wrigley et al., 2012).

Construct validity checks if all facets of the construct are included. Threats to construct validity occur if the construct is underrepresented by (1) too few questions (Downing, 2002; Downing & Haladyna, 2004). Additionally, construct validity is endangered if (2) construct-irrelevant variance occurs in the form of participants who differ in their motivation to perform (Haladyna & Downing, 2004) by showing different levels of test-taking effort. In this case test scores reflect a combination of ability and test-taking effort (Barry et al., 2010; Wise & DeMars, 2005). This problem is mentioned in combination with low-stakes assessments (Hosch, 2012; Setzer et al., 2013; Wise & DeMars, 2005; Wise, 2009). Since progress tests measure graduate knowledge, (3) percentages of correct answers should increase over the course of study ("growth of knowledge").

To investigate *criterion validity*, external criteria are used to show if there is a relation between the assessment instrument and another instrument that is intended to measure the same construct. With (1) *concurrent validity* the external instrument is measured at the same time, with (2) *predictive validity*, in the future. In knowledge assessments other knowledge tests can be used, like end-of-the-block assessments or the results of licensing examinations. Great care is needed when choosing the external assessment because it should measure the same construct. For example, knowledge assessments are less related to workplace-based assessments than to other knowledge assessments.

Validity of Progress Tests

Content Validity of Progress Tests Blueprints (1) and the use of review committees to ensure the appropriate level and content of the questions (2) are reported for moderate-stakes progress tests (Aarts et al., 2010; Lillis et al., 2014; Schuwirth et al., 2010); both are reported for low-stakes progress tests, too (Findyartini et al., 2015; Nouns & Brauns, 2008).

Construct Validity of Progress Tests The requisite range of questions (1) to prevent under-representation is fulfilled in moderate-stakes progress tests with multiple-choice or true/false questions ranging from 125 (Freeman & Ricketts, 2010; Wrigley et al., 2012) to 250 questions (van Berkel, 1990) and in low-stakes progress tests with 120 (Findyartini et al., 2015) to 200 questions (Nouns & Georg, 2010; Osterberg et al., 2006). All cover the wide range of topics of graduate medical knowledge. In the very beginning there was a progress test of 400 questions without pass/fail decisions (Willoughby, Dimond, & Smull, 1977) which later had moderate-stakes (Arnold & Willoughby, 1990) and is presumably no longer performed (Freeman et al., 2010). In low-stakes assessments there is no extrinsic motivation (Ryan & Deci, 2000) in the form of grades to *"perform at his/her best"* (Wise & DeMars, 2005, p. 2). If there are no consequences of performance, not all students will take the test seriously, the so-called non-serious test-takers, and show low test-taking effort. In low-stakes assessments, there can be a high variation in test-taking effort. Construct-irrelevant variance in the form of students with low test-taking effort (2) is

described in low-stakes progress tests (Brauns, 2007; Nouns & Georg, 2010; Osterberg et al., 2006; Schüttpelz-Brauns, 2017; Tomic et al., 2005). Osterberg et al. (2006) report 10 to 25% of non-serious participants for their low-stakes progress test. Knowledge growth (3) could be shown in many examples from moderate-stakes progress tests (Blake et al., 1996; Freeman & Ricketts, 2010; Lillis et al., 2014; Muijtjens et al., 2008; Verhoeven et al., 2002; van der Vleuten et al., 1996, 2004) as well as from low-stakes progress tests (Findyartini et al., 2015; Nouns & Georg, 2010; Osterberg et al., 2006; Schmidmaier et al., 2010; Tomic et al., 2005; Willoughby & Hutcheson, 1978), with one exception. In the pre-clinical phase of the traditional German undergraduate medical training without patient contact but with emphasis on biomedical science, no increasing test scores on a low-stakes progress test were found. In comparison to this, students from a reformed curriculum with early patient contact did have increasing test scores (Osterberg et al., 2006). It was assumed, and later found in evaluation comments, that students in the pre-clinical phase, in which basic sciences were taught independently from patients, were not able to integrate their knowledge of patients to answer the vignette questions (Osterberg et al., 2006). Therefore it can be assumed that the lack of increasing test scores is not specific to low-stakes assessment but to both if early patient contact is missing in the undergraduate training.

Criterion Validity of Progress Tests *Concurrent validity* for moderate-stakes progress tests could be shown in studies that correlated progress test results with results of practical assessments like Objective Structured Clinical Examinations (OSCE) or Mini Clinical Examinations (Mini-CEX) (Lillis et al., 2014) and in studies that used results of clinical reasoning tests (Boshuizen et al., 1997) or results of nationally standardized tests (Arnold & Willoughby, 1990) as external criteria (1). One study used ratings of students' knowledge and found high correlations with progress test results (Arnold & Willoughby, 1990). Concurrent validity could be shown as well in low-stakes progress tests. Progress test results were then correlated with the results of (moderate- and high-stakes) summative assessments (Given et al., 2016; Schmidmaier et al., 2010) and with grade point averages (Findyartini et al., 2015). Studies on *predictive validity* of moderate-stakes progress

tests (2) have shown correlations with licensure examinations for more advanced semester levels (Blake et al., 1995, 1996; Kerfoot et al., 2011). Kerfoot et al. (2011) found a positive predictive value of 41% for identifying poorly performing students in Step 1 of the United States Medical Licensing Examination (USMLE). Several studies were conducted to show predictive validity in low-stakes progress tests. Johnson et al. (2014) found correlations of progress test results and results of licensure examinations for more advanced semester levels, but it is not clear whether the progress test in this study was implemented in the curriculum or just used for research purposes. Correlations of results on low-stakes progress tests and results of licensure examinations were found in some more studies, like Willoughby et al. (1977). Karay and Schauber (2018) found that medical students' progress test growth curves were positively related to the performance on the national licensing examination. Schmidmaier et al. (2010) found a moderate correlation of progress test results and the first part of the German national licensing examination after the pre-clinical phase. Nouns et al. (2004) found high correlations between results on a low-stakes progress test and the multiple-choice sections of both parts of the national licensing examinations after the pre-clinical and clinical phases.

2.2.4 Educational Impact

Theoretical Considerations

Progress tests are meant to have educational impact by (1) identifying strengths and weaknesses in medical knowledge at the current level of undergraduate training and thus help (2) guide the future learning of the students (Norcini et al., 2018; Nouns & Georg, 2010; Schuwirth & van der Vleuten, 2012; van der Vleuten et al., 2018; Wrigley et al., 2012). Furthermore, the assessment format influences the learning styles. For example, assessments with multiple-choice questions on the factual knowledge level provoke superficial learning strategies (Cobb, Brown, Jaarsma, & Hammond, 2013; Leung, Mok, & Wong, 2008; Scouller, 1998). Because of the repeated integrated assessment of graduate-level knowledge, preparation for the progress test is impossible. Therefore progress testing should influence learning styles by discouraging binge learning (Schuwirth & van der

Vleuten, 2012) and (3) encourage learning strategies that lead to deeper and more integrated medical knowledge.

Educational Impact of Progress Tests

There were some attempts to find out if students use the results of moderate progress tests to (1) identify their strengths and weaknesses to (2) guide their future learning. In the beginning of the progress test at McMaster's University, Blake et al. (1995) asked students about their experience with the progress test. The students stated that the progress test results were only slightly helpful for identifying strengths and weaknesses. One year later Blake et al. (1996) found that progress test results played a moderate role in identifying strengths and weaknesses. Aarts et al. (2010) found in interviews that the majority of students use their results to monitor their knowledge growth but they did not mention the impact on learning itself. Wade et al. (2012) conducted a questionnaire in two different settings and found that students from the school with feedback on progress test performance by discipline agreed more that the progress test helped to improve their knowledge and to monitor the improvement, but there was no hint that the results of the test guided their learning. In Auckland focus groups discussed the impact on learning. Subsequent analyses showed that junior students' future learning was guided by the content of the progress test itself, and not by the feedback. For senior students the progress test was used to bring students back to the core learning and reinforce it (Yielder et al., 2017). Similarly, Given et al. (2016) found in semi-structured interviews that students who took a low-stakes progress test felt informed about their strengths and weaknesses, but the feedback did not guide their future learning. Nouns and Georg (2010) reported that students used their results on the low-stakes progress test for their learning, but unfortunately there was no mention of how. Regarding the impact of progress testing on learning styles (3), there are only publications from working groups focused on moderate-stakes progress tests. The results are contrary. On the one hand, there was no evidence in surveys with self-developed, as well as validated questionnaires, that students changed their conceptual learning strategy due to the introduction of a

progress test (Blake et al., 1995, 1996; Chen et al., 2015). On the other hand, a moderate-stakes progress test encouraged deeper and more integrated medical knowledge, which Lillis et al. (2014) found when asking students to fill out a self-developed questionnaire. In a study van Berkel et al. (1995) asked students about their learning strategies and their results in progress testing using a validated questionnaire and found that students who learned in a meaning-oriented manner had higher scores in progress testing than students who memorized knowledge. Schuwirth and van der Vleuten (2012) report that students changed their learning strategies after the introduction of the moderate-stakes progress test together with moderate-stakes block tests. After the stakes of the block tests were raised, students went back to short-term memorization learning strategies. There is no published study on the impact on learning strategies in low-stakes progress tests.

2.2.5 Acceptability

Theoretical Considerations

Assessments and assessment results should be accepted by those affected, in particular faculty and students (Verhoeven, 2003). Otherwise, the assessments are not taken seriously (Verhoeven, 2003) and will not last (van der Vleuten, 1996; van der Vleuten et al., 2000).

Acceptability of Progress Tests

In regard to moderate-stakes progress tests, students report a positive attitude towards the test as a useful assessment that supports their learning (Ali, Cockerill, Zahra, Tredwin, & Ferguson, 2018). But there are also two groups reporting resistance (Aarts et al., 2010) and high running feelings (Blake et al., 1995) after introducing a moderate-stakes progress test because faculty and students lacked experience with this formerly unknown type of testing. It is not reported how the resistance manifested itself. Acceptance increased over the time at both medical schools. Aarts et al. (2010) reported that students got used to the concept and accepted the test more when they gained awareness of their knowledge growth. Tomic et al. (2005) reported that resistance increased after a low-

stakes progress test became compulsory. Students registered their attendance, but did not answer any questions on the test. Similarly, Osterberg et al. (2006) reported 10% to 25% non-serious participation in another low-stakes progress test. In 2009 results of a survey among participants of a low-stakes progress test showed that students accepted the test more if they understood the concept of formative assessment and used the results for their learning (Nouns & Georg, 2010). It was not reported how many students changed their acceptance or how many students used the results for their learning.

2.2.6 Cost-Effectiveness

Theoretical Considerations

Progress tests are very cost-intensive (Albanese & Case, 2016) if the faculty has an elaborate quality management system to write, review, and manage questions that fulfill the demands of a progress test, to create and administer the test, to analyze data and create feedback reports for students, faculty, question authors, etc. Although costs may be high, it is worthwhile to invest in high-quality progress tests to ensure high reliability, validity, educational impact and therefore acceptance by students and faculty.

In collaborations the cost of developing the progress tests and psychometric expertise is shared (Albanese & Case, 2016; Findyartini et al., 2015; van der Vleuten, Schuwirth, Scheele, Driessen, & Hodges, 2010; van der Vleuten et al., 2018; Wrigley et al., 2012). This reduces the costs (Schuwirth et al., 2010; Schuwirth & van der Vleuten, 2012), represents a useful quality assurance tool (Finucane et al., 2010) and allows for benchmarking (Schuwirth et al., 2010; Schuwirth & van der Vleuten, 2012).

Cost-Effectiveness of Progress Tests

Collaborations that can be found on the internet develop moderate-stakes progress tests like the Interuniversity Progress Test Medicine involving five universities from the Netherlands (http://ivtg.nl) or low-stakes progress tests like the Berlin Progress Test with more than 15 universities from German-speaking countries (https://progress-test-medizin.charite.de/en/). In the International Partnership for Progress Testing (http://ipptx.org), universities from

Portugal, Ireland, and Canada can decide whether to use the progress test with moderate or high stakes. On the homepage of the EBMA International Progress Test (www.ebma.eu/ipt), several participating countries, such as the Netherlands, Mexico, Australia, Mozambique, and Saudi Arabia, are listed, but no further information on the stakes is given.

In Freeman et al. (2010), collaborations between the National Board of Examiners (NBME) and medical schools in the United Kingdom, as well as with a U.S. medical school, are described. The NBME currently offers knowledge tests commercially (https://www.nbme.org/Students/sas/sas.html).

2.2.7 Summarizing the Results

Table 2.1 summarizes of the results of the literature review.

As can be seen, low-stakes progress tests are comparable regarding the utility components with two exceptions:

- construct-irrelevant variance due to students with low test-taking effort, and
- acceptance by students and faculty.

There are two questions that are not definitely answered yet for both moderate- and low-stakes progress tests concerning the educational impact: how or under which conditions do progress test results guide future learning, and how do progress tests encourage deeper learning strategies. These two questions are not considered in this work.

Table 2.1: **Utility of moderate- and low-stakes progress tests: Results of a literature review**

Utility Component	Criteria	Moderate-stakes pt	Low-stakes pt
Reliability	(1) large number of questions	✓	✓
	(2) formal requirements for questions	✓	✓
	(3) requirements for level of reliability		
	- entire sample	✓	✓
	- within cohorts	×	✓
Validity			
- Content	(1) blueprint	✓	✓
	(2) content requirements for questions	✓	✓
- Construct	(1) wide range of questions	✓	✓
	(2) no construct-irrelevant variance by low test-taking effort	✓	×
	(3) growth of knowledge over the study years*	✓	✓
- Criterion			
– Concurrent	(1) correlation with contemporaneous assessment	✓	✓
– Predictive	(2) correlation with future assessment	✓	✓
Educational impact			
	(1) identify strengths and weaknesses	✓	✓
	(2) guide future learning	?	?
	(3) encourage deeper learning strategies	?	?
Acceptability	accepted by students and faculty	✓	×
Costs	collaborations to reduce costs	✓	✓

Note. * only if there is early patient contact in the undergraduate training, ✓requirements are met, × requirements are not met, ? not answered yet

Chapter 3

Research Aims

As with other low-stakes assessments, special considerations have to be taken in low-stakes progress tests (1) regarding threats to the construct validity due to high variations in test-taking effort (Attali, 2016; Barry et al., 2010; Butler & Adams, 2007; Eklöf, 2010; Levine & Rubin, 1979; Setzer et al., 2013; Schmitt, Chan, Sacco, McFarland, & Jennings, 1999; Waskiewicz, 2011; Wise & DeMars, 2010) and (2) regarding their acceptability.

There are two common ways to deal with the problem of decreased construct validity due to high variations in test-taking effort. First is to make participation voluntary. In this case participating students should have high test-taking effort. However, experience with the low-stakes Berlin Progress Test shows that this produces highly selective samples. On average, 8% (range: 0% to 70%) of a semester were taking part when attendance was voluntary at one medical school, at another medical school an average of 4% with a range from 0% to 12% (unpublished data). Tomic et al. (2005) reported that attendance of a voluntary low-stakes progress test was on average 28% (range: 6%-65% per semester). A highly selective sample in a voluntary progress test strongly risks distorting the results in a positive way and therefore can affect the validity of the results.

In the second approach, students with low test-taking effort, so-called non-serious test-takers, are identified and excluded from further analyses (Wise & DeMars, 2005). This can be done over the processing time (Brauns, 2007; Wise & Kong, 2005). This is an objective marker (Finn, 2015; Wise & Kong, 2005). Computer-based it is reliable. The problem is the cut-off value which can lead to a false identification of “fast” geniuses

(Brauns, 2007). In the person-fit indices non-serious test-takers are identified on the basis of statistical models (Brauns, 2007; Meijer, 1996; Meijer & Sijtsma, 2001). The advantage are the underlying, empirically tested theoretical models (Brauns, 2007). The detection rates can be relatively high, especially with long tests (Nering & Meijer, 1998) or they can also be low to medium (Drasgow, Levine, & McLaughlin, 1987). Especially with high test values, the specificity can be very low (Brauns, 2007). Questionnaires are also used to identify test-taking effort in participants (Crombach, Boekaerts, & Voeten, 2003; Rheinberg, Vollmeyer, & Burns, 2001; Sundre & Moore, 2002). Questionnaires are relatively easy to collect and evaluate (Swerdzewski, Harmes, & Finney, 2011). They can be administered both paper-based and computer-based. Until now, there has been no questionnaire with a cut-off value for identifying non-serious test-takers (Finn, 2015), plus they were rather long, with one exception (see Effort Thermometer in Baumert and Demmrich (2001)). The disadvantage of questionnaires is that they are not immune to socially desirable answers (Wise & Kong, 2005; Wise & Ma, 2012), or to a low effort while filling out those questionnaires (Finn, 2015). This disadvantage could be alleviated by a short scale.

This leads to the first research aim of *developing and validating a short scale for identifying students with low test-taking effort.*

If the developed instrument can correctly identify non-serious students in low-stakes progress tests, construct-irrelevant variance due to non-serious test-takers is decreased and the test can, for example, be used as a research tool.

This leads to the second research aim of *investigating the construct validity of a low-stakes progress test* after eliminating non-serious test-takers.

There is a third way to deal with the problem of non-serious test-takers on low-stakes progress tests: increasing the test-taking motivation (Finn, 2015; Wise & DeMars, 2005) and thus the acceptability, the second utility component which needs special consideration in terms of low-stakes progress tests, in contrast to moderate-stakes progress tests.

This leads to the third research aim of finding *strategies that are related to the acceptability* of low-stakes progress tests.

These three research aims are now described in more detail.

3.1 Developing and Validating a Short Scale for Identifying Students with Low Test-Taking Effort

Due to non-serious test-takers, construct-irrelevant variance increases within low-stakes progress tests. One way to handle this is to identify non-serious test-takers and remove their results from further analysis. Non-serious test-takers on the BPT are identified with a combination of various objective criteria (Nouns & Georg, 2010). However, some of these require statistical expertise. Therefore, we looked for a way to easily identify non-serious test-takers of the low-stakes progress test by measuring test-taking effort in a cheap and easy way without needing any statistical expertise.

Therefore, we have developed a short questionnaire to measure test-taking effort based on expectancy-value theory (Wigfield & Eccles, 2000) and conducted a validation study.

The aim of this study was (1) to develop a short test-effort self-assessment scale that is capable of measuring test-taking effort in low-stakes testing with high reliability and validity and (2) to conduct a validation study for the scale developed.

3.2 Investigating the Construct Validity of a Low-Stakes Progress Test

If non-serious test-takers are identified and excluded from further analyses, construct-irrelevant variance should decrease and construct validity should increase.

We therefore conducted a study which investigated if differences in two curricula regarding the acquisition of knowledge can be reconstructed with the help of a low-stakes progress test.

The aim of this study was to compare the development and retention of knowledge

in the basic medical sciences between students enrolled in the traditional and reformed undergraduate medical curricula.

3.3 Strategies Related to the Acceptability of a Low-Stakes Progress Test

When introducing progress tests there was resistance to both moderate- and low-stakes variants, which, after a while, receded for moderate-stakes progress tests (Aarts et al., 2010; Blake et al., 1996). For the low-stakes test, the problem of lack of acceptance is permanent and therefore needs special attention. Successful strategies to increase acceptance should reduce the proportion of non-serious test-takers.

Therefore, we conducted several studies to examine strategies that increase the acceptability of a low-stakes progress test.

3.3.1 Introducing Computer-Based Assessment to Increase Acceptability of a Low-Stakes Progress Test

One of the strategies can be to provide feedback immediately (Irons, 2008; Shute, 2008). This can be done by changing from paper-based to computer-based administration of the test, since the evaluation of paper-based tests takes several days or weeks, while the first results of computer-based tests are already available during the test.

We conducted a study in which we compared the acceptance of a low-stakes progress test prior to and after introduction of computer-based administration.

The aim of this study was to show whether immediate feedback by introducing computer-based administration would increase the students' acceptance of a low-stakes progress test.

3.3.2 Effects of Changing from Paper-Based to Computer-Based Test Format

When changing the format from paper- to computer-based, it must be ensured that performance is not influenced by the test format. This we tested in another study with a randomized matched-pair design.

The aim of this study was to investigate whether computer-based tests influence students' test performance by comparing the test performance of students taking the paper-based and the computer-based versions of the same low-stakes progress test.

3.3.3 Institutional Strategies Related to Acceptability of a Low-Stakes Progress Test

Furthermore, there will be other strategies that can potentially increase the test-taking effort on low-stakes assessments. These strategies can be derived from motivational theories.

In another study we derived potential strategies from self-determination theory (Ryan & Deci, 2000) that medical schools can carry out to increase acceptability of their low-stakes progress test.

In this study we aimed to identify institutional factors related to test-taking effort in a low-stakes progress test to provide medical schools with practical recommendations on how to increase its utility.

Chapter 4

Empirical studies

4.1 Developing and Validating a Short Scale for Identifying Students with Low Test-Taking Effort

Schüttpelz-Brauns, K., Kadmon, M., Kiessling, C., Karay, Y., Gestmann, M. & Kämmer, J.E. (2018). Identifying low test-taking effort during low-stakes tests with the new Test-taking Effort Short Scale (TESS) - Development and Psychometrics. *BMC Medical Education, 18(1)*, 101.

4.2 Investigating the Construct Validity of a Low-Stakes Progress Test

Nouns, Z.M., Schauber, S., Witt, C., Kingreen, H. & Schüttpelz-Brauns, K. (2012) Development of knowledge in basic medical sciences during undergraduate medical education – a comparison of a traditional and a problem-based curriculum. *Medical Education, 46*(12): 1206-1214.

4.3 Strategies Related to Acceptability of Low-Stakes Progress Tests

4.3.1 Introducing Computer-Based Assessment to Increase Acceptability of a Low-Stakes Progress Test

Karay, Y., Schauber, S., Stosch, C. & Schuettpelz-Brauns, K. (2012) Can computer-based assessment enhance the acceptance of formative multiple choice exams? A utility analysis. *Medical Teacher, 34*(4): 292–296.

4.3.2 Effects of Changing from Paper-Based to Computer-Based Testformat

Karay, Y., Schauber, S.K., Stosch, C. & Schüttpelz-Brauns, K. (2015). Computer versus paper – does it make any difference in test performance? *Teaching and Learning in Medicine, 27*(1), 57-62.

4.3.3 Institutional Strategies Related to the Acceptability of a Low-Stakes Progress Test

Schüttpelz-Brauns, K., Hecht, M., Hardt, K., Karay, Y., Zupanic, M. & Kämmer, J.E. (2019) Institutional strategies related to test-taking behavior in low stakes assessment. *Advances in Health Sciences Education, 25*, 331-335.

Chapter 5

Discussion

In my work I compared previous findings for moderate-stakes and low-stakes progress tests regarding the utility components of assessment (van der Vleuten, 1996). Low-stakes progress tests do not have negative consequences if students fail to perform at their best (Wise & DeMars, 2005). This leads to high variations in test-taking effort. Performance then not only depends on ability, but also on test-taking effort (Barry et al., 2010; Eklöf & Knekta, 2017; O'Neil et al., 1995; Sundre & Kitsantas, 2004; Wise & Kong, 2005; Wolf & Smith, 1995) with consequences for the validity of test results. The acceptability of low-stakes progress tests can therefore be lower than for moderate-stakes progress tests. This led to the questions of (1) how to identify non-serious test-takers and if - after eliminating non-serious test-takers - (2) construct validity of a low-stakes progress test can be shown. Another aim was to (3) identify and implement strategies that increase the acceptability of the low-stakes progress test, which we investigated in three further studies.

In our first study we *developed and validated a short scale for identifying students with low test-taking effort*. We could show very good psychometrics including construct validity of the three-item Test-taking Effort Short Scale (TESS). We were able to develop and validate a scale shorter than previously published scales for measuring test motivation (Crombach et al., 2003; Rheinberg et al., 2001; Sundre & Moore, 2002; Thelk et al., 2009). There is only one shorter scale, the Test Effort Thermometer (Baumert & Demmrich, 2001), but no published studies on its psychometrics and, in contrast to TESS, it is not based on a theoretical framework. TESS is the first questionnaire that measures

test-taking effort and provides a cut-off value for identifying non-serious test-takers. To increase construct validity of the low-stakes Berlin Progress Test, non-serious test-takers were already identified with appropriateness measurement (Brauns, 2007; Nouns & Georg, 2010; Osterberg et al., 2006) and were eliminated from the analyses. However, there is another advantage to TESS over prior statistical measures. By using a questionnaire rather than response time error (Brauns, 2007; Wise & Kong, 2005) or appropriateness measurement (Brauns, 2007; Meijer, Muijtjens, & van der Vleuten, 1996; Meijer & Sijtsma, 2001; Meijer, 2003) to statistically identify test-takers with low test-taking effort, participating students see that administrators are concerned about the problem of low test-taking effort. In our experience, students with average test-taking effort are likely to increase their effort if they know that the resulting individual feedback of a low-stakes assessment will not be negatively influenced by non-serious test-takers.

In the second study we *investigated the construct validity of a low-stakes progress test.* After eliminating non-serious test-takers, the effects of teaching hours and learning for a high-stakes assessment on the number of correct answers on the low-stakes Berlin Progress Test could be shown by comparing students in traditional versus reformed undergraduate medical curricula at the same medical school. We did not investigate whether we could show the difference between both curricula if non-serious test-takers were not excluded from analysis. There are mixed results for the impact of non-serious test-takers on results in low-stakes assessments. Some studies found that test-taking effort had an obvious effect on performance in low-stakes assessments (Akyol et al., 2018; Brown & Walberg, 1993; Eklöf, 2010; Wise & DeMars, 2005; Wolf & Smith, 1995). Because consequences for schools and policy from low-stakes assessments can be severe (Breakspear, 2012), it is important to further study the impact of non-serious test-takers on the validity of results in large scale assessments.

In the third to sixth studies we investigated *strategies meant to be related to acceptability of a low-stakes progress test.*

In the third study by *introducing computer-based assessment* we showed that immediate feedback *increased acceptability of a low-stakes progress test.* We found small effects

on the proportion of positive comments in the evaluation as well as on the proportion of serious participants in favor of the computer-based administration. Hence, the acceptability and thus the overall utility of low-stakes progress tests can be enhanced by providing immediate feedback via computer-based assessment. This finding is consistent with earlier research on formative assessment and feedback in the field of psychology (Csikszentmihalyi & Lefevre, 1989; Kulik & Kulik, 1988; Skinner, 1958; Tuten, Galesic, & Bosnjak, 2004) and also with recent research in dental medicine (Zheng & Bender, 2018).

In the fourth study we showed that *changing from the paper-based to computer-based test format* had no effect on students' performance on the low-stakes Berlin Progress Test within a randomized matched-pair design. This is in accordance with the study of Hochlehnert, Brass, Moeltner, and Jünger (2011) which investigated self-selected groups. This means computer-based low-stakes assessment can be used as formative assessment without disadvantaging the participants.

In the fifth study we identified *institutional strategies related to acceptability of a low-stakes progress test* which were drawn from the self-determination theory of Ryan and Deci (2000). We found connections with the following strategies: (1) discussion of low performance with the mentor, (2) consequences for not participating, (3) give choice of place and date of test taking. Serious test-taking behavior was more likely if students were given choices *and* the low-stakes progress test was presented as assessment, or students were given no choices *and* the test was presented as evaluation. There are several authors who emphasize the importance of providing dialogue for effective feedback (Irons, 2008; Nicol & Macfarlane-Dick, 2006; Smyth, 2004; van der Vleuten, Schuwirth, Driessen, Govaerts, & Heeneman, 2015) which is fulfilled in strategy (1). When discussing low performance in progress tests, a mentor can provide a safe environment and a teacher/learner relationship, which is important when using assessment to support learning (Schut, Driessen, van Tartwijk, van der Vleuten, & Heeneman, 2018). Mentors can facilitate or hinder receptivity to feedback (Harrison et al., 2016). Therefore it requires a commitment by faculty, especially mentors, to offer supportive mentorship and *"create a learning environ-*

ment free from the restraints of traditional assessment design" (Harrison & Wass, 2016, p. 705). Although studies have shown that consequences (2) like grading (Baumert & Demmrich, 2001) or consequences for the institution (Liu, Rios, & Borden, 2015) impact performance, there is, to my knowledge, no published study showing that consequences of not participating also have an effect of the performance on low-stakes assessments. Giving students choice (3) can enhance the intrinsic motivation and thus performance (Patall, Cooper, & Robinson, 2008). In contrast to these prior findings of the meta-analysis, we found hints that the motivation level (extrinsic vs. intrinsic level) might have an influence on the strategy for giving choices for students. Not all strategies we derived from self-determination theory (Ryan & Deci, 2000) showed the expected effects. An assessment system within a curriculum is such a complex system with many (unkonwn) influencing variables that large effects are needed to statistically prove an influence (Ringsted, Hodges, & Scherpbier, 2011). On the other hand, some strategies might be good in theory, but don't work in practice as intended. In this study there are some questions that remain open. Firstly, there are no studies that examine the relationship between the effectiveness of institutional strategies and the motivation levels of students. This would take into account the perspective of the students. In this context, it would also be important to examine the influence of peers and the influence of the social and emotional support from teachers on commitment to learning. There are some studies from the school sector that provide evidence that these factors have an influence and must therefore be taken into account in the context of investigations of test efforts in low-stakes assessment (Ketonen & Hotulainen, 2019; Kindermann, 2007; Ruzek et al., 2016; Warburton, 2017; Wentzel, Battle, Russell, & Looney, 2010; Wentzel, Muenks, McNeisha, & Russell, 2017).

Our studies stand out from the literature because we were able to study large amounts of data and long periods of time, and we were able to include different medical schools with different curricula and implementation conditions for the low-stakes Berlin Progress Test. In our research we could show that there are several options to identify students with low test-taking effort. Low-stakes assessments can be made more valid by eliminating their results from further analyses (Barry et al., 2010; Hosch, 2012; Setzer et al., 2013;

Wise, 2009; Wise & DeMars, 2005). This is especially important in the case of large-scale assessments, which are low-stakes for students, but high-stakes for institutions and policy (Breakspear, 2012; Fullan, 2009).

On the other hand, we identified strategies to increase acceptability of low-stakes assessment which is especially important for low-stakes assessments that are used as feedback instruments, e.g. in self-regulated learning settings. These strategies are

- computer-based assessment with immediate feedback,
- discussion of low performance with a mentor,
- consequences for not participating,
- if you give students a choice of place and time for taking a low-stakes progress test, integrate the low-stakes test into the assessment system rather than into the evaluation system.

There are several limitations of the research that have to be mentioned. The restrictions of each study are discussed within each paper. The *literature review* conducted in the background chapter that led to the research aims also has limitations. First, there are language restrictions: some publications on progress tests might be in the language of the university where they are administered. I only used information that was available in English or German making it possible that the information presented in the background chapter is not complete. Of the low-stakes progress tests presented in the background chapter, the progress test in Finland (Freeman et al., 2010) is not included in the literature review due to a lack of publications in English. Second, some of the considered publications are rather old. It is conceivable that some of the progress tests no longer exist or have changed. For example, the progress test at the University of Missouri-Kansas City had low-stakes consequences in the beginning (Willoughby et al., 1977) and moderate consequences some years later (Arnold & Willoughby, 1990). This should not affect the conclusions from the comparison because the papers referred to the respective versions of either the low-stakes or the moderate-stakes progress test. Third, there is a limitation of my literature research that affects the problem of non-serious test-takers.

It is possible that the problem of non-serious test-takers is not published on by each working group where this problem occurs. Thus, we probably face a classic file-drawer problem (Dickersin, 1990; Scargle, 2000). There might be an unknown number of studies or project reports that were not published due to non-serious test-takers in low-stakes progress tests. On the other hand, there might be an unknown number of non-serious test-takers in moderate-stakes progress testing as well. In moderate stakes progress tests, students' results are accumulated over a certain number of successive tests. It might be conceivable that students who "passed" the required number of successive tests will have less test-taking effort on the final of successive test(s) and therefore their performance would decrease; however, there is no published study on this issue. Therefore it is possible that the systematic comparison of moderate-stakes and low-stakes progress tests is biased. Based on the long experience with the low-stakes Berlin Progress Test, the challenges of low-stakes progress tests are known and possible biases can be discerned, but not for moderate-stakes progress tests. Another limitation is the *number of controlled or uncontrolled variables* which are taken into account in one study. Many influencing variables play a role when considering the strategies for increasing the acceptability of a low-stakes test. Results from studies at one faculty in which influencing factors are kept constant are probably less generalizable to other faculties than multisite studies. In multisite studies with many uncontrollable influences, however, the measured effects become smaller. In this work I only considered *low-stakes knowledge progress tests with large samples of test-takers.* I excluded all other kinds of low-stakes assessment, like formative workplace-based assessment. In situations where one facilitator observes one student and gives feedback it is unlikely that students don't give their best. Additionally, in the studies of acceptability we focused on strategies that concentrate on the individual level of the low-stakes progress test, meaning the purpose of formative assessment. Therefore all findings in my acceptability studies only help to improve low-stakes assessment with formative purpose and large samples.

One question on the individual (feedback) level remains open: the educational impact of "assessments for learning". It is still unclear under which conditions low-stakes and

moderate-stakes assessments serve the purpose of guiding the learning process, which is part of self-regulated learning. Studying the functioning of low-stakes progress tests can be a great chance to understand the process of self-regulated learning in different settings. One might argue that it is too much effort to increase acceptability of a low-stakes assessment and therefore stakes have to be raised to solve the problem of non-serious test-takers. If the stakes are raised, acceptability might increase (Wolf & Smith, 1995), but also test anxiety (Cassady & Johnson, 2002), costs of administration and required resources (Brown & Finney, 2011; Wise & DeMars, 2005). But the most important focus on "assessment for learning" will shift back to "assessment of learning". A lack of educational impact, more precisely the impact on future learning in formative assessment, will be the consequence (Hawthorne, Bol, Pribesh, & Suh, 2015). If we simply raise stakes, there might be the danger of thinking anything goes because there are no non-serious test-takers, but they are only the symptom of a non-functioning formative assessment. If we do not question the utility of our assessments, it is, without any doubt, easy to run a non-functioning assessment. Only when we do understand the influencing variables of utility, especially test-taking effort and acceptability, in low-stakes assessments and use low-stakes (formative) assessment as a comprehensible part of undergraduate training will we have students who focus on learning for life, rather than learning how to effectively pass assessments. Grades are external motivators. We should instead focus on the learning process and on intrinsic motivators for learning because the learning process lasts throughout life and external motivators are not always available. As long as we see assessments as end-of-whatever measurements, they will only be useful if they motivate extrinsically and students will be grade-driven instead of self-regulated.

References

Aarts, R., Steidel, K., Manuel, B., & Driessen, E. (2010). Progress testing in resource-poor countries: A case from Mozambique. *Medical Teacher*, *32*(6), 461-463. doi: 10.3109/0142159X.2010.486059

Akyol, S., Krishna, K., & Wang, J. (2018). *Taking PISA seriously: how accurate are low stakes exams?* (NBER Working Paper No. No. 24930). National Bureau of Ecomonic Research. doi: 10.3386/w24930

Al Alwan, I., Al-Moamary, M., Al-Attas, N., Al Kushi, A., AlBanyan, E., Zamakhshary, M., ... Schmidt, H. (2011). The progress test as a diagnostic tool for a new PBL curriculum. *Education for Health (Abingdon)*, *24*(3), 493.

Albanese, M., & Case, S. (2016). Progress testing: critical analysis and suggested practices. *Advances in Health Sciences Education*, *21*(1), 221-234. doi: 10.1007/s10459-015-9587-z

Ali, K., Cockerill, J., Zahra, D., Tredwin, C., & Ferguson, C. (2018). Impact of progress testing on the learning experiences of students in medicine, dentistry and dental therapy. *BMC Medical Education*, *18*(1), 253. doi: 10.1186/s12909-018-1357-1

Ali, K., Coombes, L., Kay, E., Tredwin, C., Jones, G., & Ricketts, J., C. and. Bennett. (2016). Progress testing in undergraduate dental education: the Peninsula experience and future opportunities. *European Journal of Dental Education*, *20*(3), 126-134. doi: https://doi.org/10.1111/eje.12149

Alvermann, D., Smith, L., & Readence, J. (1985). Prior knowledge activation and the comprehension of compatible and incompatible text. *Reading Research Quarterly*, *20*, 420-436. doi: 10.2307/747852

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1999). *Standards for education and psychological testing.* Washington, DC: American Educational Research Association.

Arnold, L., & Willoughby, T. (1990). The quarterly profile examination. *Academic Medicine*, *65*(8), 515-516. doi: https://doi.org/10.1177/001316447803800425

Attali, Y. (2016). Effort in low-stakes assessments: what does it take to perform as well as in high-stakes setting? *Educational and Psychological Measurement*, *76*(6), 1045-1058. doi: https://doi.org/10.1177/0013164416634789

Bae, C., Therriault, D., & Redifer, J. (2018). Investigating the testing effect: retrieval as a characteristic of effective study strategies. *Learning and Instruction*. doi: https://doi.org/10.1016/j.learninstruc.2017.12.008

Bangert-Drowns, R., Kulik, J., & Kulik, C. (1991). Effects of frequent classroom testing. *The Journal of Educational Research*, *85*(2), 89-99. doi: https://doi.org/10.1080/00220671.1991.10702818

Barry, C., Horst, S., Finney, S., Brown, A., & Kopp, J. (2010). Do examinees have similar test-taking effort? A high-stakes question for low-stakes testing. *International Journal of Testing*, *10*(4), 342-363. doi: https://doi.org/10.1080/15305058.2010.508569

Baumert, J., & Demmrich, A. (2001). Test motivation in the assessment of student skills: the effects of incentives on motivation and performance. *European Journal of Psychology of Education*, *16*(3), 441-462. doi: https://doi.org/10.1007/BF03173192

Bell, D., Harless, C., Higa, J., Bjork, E., Bjork, R., Bazargan, M., & Mangione, C. (2008). Knowledge retention after an online tutorial: a randomized educational experiment among resident physicians. *Journal of General Internal Medicine*, *23*(8), 1164-1171. doi: https://doi.org/10.1007/s11606-008-0604-2

Bennett, J., Freeman, A., Coombes, L., Kay, L., & Ricketts, C. (2010). Adaptation of medical progress testing to a dental setting. *Medical Teacher*, *32*(6), 500-502. doi: 10.3109/0142159X.2010.486057

Bianchi, F., Stobbe, K., & Eva, K. (2008). Comparing academic performance of medical students in distributed learning sites: the McMaster experience. *Medical Teacher*, *30*(1), 67-71. doi: 10.1080/01421590701754144

Black, P., & William, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy and Practice*, *5*, 7-74. doi: https://doi.org/10.1080/0969595980050102

Blake, J., Norman, G., Keane, D., Mueller, C., Cunnington, J., & Didyk, N. (1996).

Introducing progress testing in McMaster University's problem-based medical curriculum: psychometric properties and effect on learning. *Academic Medicine*, *71*(9), 1002-1007.

Blake, J., Norman, G., & Smith, E. (1995). Report card from McMaster: student evaluation at a problem-based medical school. *The Lancet*, *345*(8954), 899-902. doi: https://doi.org/10.1016/S0140-6736(95)90014-4

Boshuizen, H., van der Vleuten, C., Schmidt, H., & Machiels-Bongaerts, M. (1997). Measuring knowledge and clinical reasoning skills in a problem-based curriculum. *Medical Education*, *31*(2), 115-121. doi: https://doi.org/10.1111/j.1365-2923.1997.tb02469.x

Brauns, K. (2007). *Identifikation von Musterkreuzern beim Progress Test Medizin* (Unpublished doctoral dissertation). Humboldt-Universität zu Berlin.

Breakspear, S. (2012). *The policy impact of PISA: An exploration of the normative effects of international benchmarking in school system performance* (OECD Education Working Papers, No. 71). OECD Education Working Papers, No. 71, OECD Publishing.

Brown, A., & Finney, S. (2011). Low-stakes testing and psychological reactance: Using the Hong Psychological Reactance Scale to better understand compliant and non-compliant examinees. *International Journal of Testing*, *11*(3), 348–270. doi: https://doi.org/10.1080/15305058.2011.570884

Brown, S., & Walberg, H. (1993). Motivational effects on test scores of elementary students. *The Journal of Educational Research*, *86*(3), 133-136. doi: https://doi.org/10.1080/00220671.1993.9941151

Brunner, M., Artelt, C., Krauss, S., & Baumert, J. (2007). Coaching for the PISA test. *Learning and Instruction*, *17*(2), 111-122. doi: https://doi.org/10.1016/j.learninstruc.2007.01.002

Butler, J., & Adams, R. (2007). The impact of differential investment of student effort on the outcome of international studies. *Journal of Applied Measurement*, *8*(3), 279-304.

Campbell, J., Voelkl, K., & Donahue, P. (1998). *NAEP 1996 Trends in Academic Progress. Achievement of U.S. Students in Science, 1969 to 1996; Mathematics, 1973 to 1996; Reading, 1971 to 1996; Writing, 1984 to 1996* (Research Report). Princeton, NJ: Educational Testing Service, Princeton.

Case, S., & Swanson, D. (2002). *Constructing written test questions for the basic and clinical sciences* (3rd revised edition ed.). Philadelphia: National Board of Medical Examiners.

Cassady, J. C., & Johnson, R. E. (2002). Cognitive test anxiety and academic performance. *Contemporary Educational Psychology*, *27*(2), 270–295. doi: https://doi.org/10.1006/ceps.2001.1094

Cecilio-Fernandes, D., Aalders, W., Bremers, A., Tio, R., & de Vries J. (2018). The impact of curriculum design in the acquisition of knowledge of oncology: Comparison among four medical schools. *Journal of Cancer Education*, *33*(5), 1110-1114. doi: 10.1007/s13187-017-1219-2

Chang, S. (2017). The effects of test trial and processing level on immediate and delayed retention. *Europe's Journal of Psychology*, *13*(1), 129–142. doi: 10.5964/ejop.v13i1.1131

Chen, Y., Henning, M., Yielder, J., Jones, R., Wearn, A., & Weller, J. (2015). Progress testing in the medical curriculum: students' approaches to learning and perceived stress. *BMC Medical Education*, *15*, 147. doi: https://doi.org/10.1186/s12909-015-0426-y

Cobb, K. A., Brown, G., Jaarsma, D. A., & Hammond, R. A. (2013). The educational impact of assessment: a comparison of DOPS and MCQs. *Medical Teacher*, *35*(11), e1598-1607. doi: https://doi.org/10.3109/0142159X.2013.803061

Coelho, C., Zahra, D., Ali, K., & Tredwin, C. (2019). To accept or decline academic remediation: What difference does it make? *Medical Teacher*, *41*(7), 824-829. doi: 10.1080/0142159X.2019.1585789

Cole, J. (2007). *Motivation to do well on low-stakes tests* (Dissertation). Faculty of the Graduate School at the University of Missouri-Columbia.

Cole, J., Bergin, D., & Whittaker, T. (2008). Predicting student achievement for low stakes tests with effort and task value. *Contemporary Educational Psychology*, *33*(4), 609-624. doi: https://doi.org/10.1016/j.cedpsych.2007.10.002

Cole, J., & Osterlind, S. (2008). Investigating differences between low- and high-stakes test performance on a general education exam. *Journal of General Education*, *57*(2), 119-130. doi: https://doi.org/10.1353/jge.0.0018

Coombes, L., Ricketts, C., Freeman, A., & Stratford, J. (2010). Beyond assessment: feedback for individuals and institutions based on the progress test. *Medical Teacher*, *32*(6), 486-490. doi: 10.3109/0142159X.2010.485652

Copp, D. T. (2018). Policy incentives in Canadian large-scale assessment: How policy levers influence teacher decisions about instructional change. *Education Policy Analysis Archives*, *25*(115), 1-39. doi: http://dx.doi.org/10.14507/epaa.25.3299

Crombach, M., Boekaerts, M., & Voeten, M. (2003). Online measurement of appraisals of students faced with curricular tasks. *Educational and Psychological Measurement*, *63*(1), 96-111. doi: https://doi.org/10.1177/0013164402239319

Crooks, N., & Alibali, M. (2013). Noticing relevant problem features: activating prior knowledge affects problem solving by guiding encoding. *Frontiers in Psychology*, *4*, 884. doi: 10.3389/fpsyg.2013.00884

Csikszentmihalyi, M., & Lefevre, J. (1989). Optimal experience in work and leisure. *Journal of Personality and Social Psychology*, *56*(5), 815–822. doi: https://doi.org/10.1037/0022-3514.56.5.815

Custers, E. (2008). Long-term retention of basic science knowledge: a review study. *Advances in Health Sciences Education*, *15*(1), 109-128. doi: https://doi.org/10.1007/s10459-008-9101-y

Custers, E., & ten Cate, O. (2011). Very long - term retention of basic science knowledge in doctors after graduation. *Medical Education*, *45*(4), 422-430. doi: https://doi.org/10.1111/j.1365-2923.2010.03889.x

Dankbaar, M., Richters, O., Kalkman, C., Prins, G., Ten Cate, O., van Merriënboer, J., & Schuit, S. (2017). Comparative effectiveness of a serious game and an e-module to

support patient safety knowledge and awareness. *BMC Medical Education*, *17*(1), 30. doi: 10.1186/s12909-016-0836-5

da Rosa, M., Isoppol, C., Cattaneol, H., Madeiral, K., Adamil, F., & Filholl, O. (2017). Progress testing as an indicator for improvements in a medical school. *Revista Brasileira de Educação Médica*, *41*(1). doi: http://dx.doi.org/10.1590/1981-52712015v41n1rb20160022

De Champlain, A., Cuddy, M., Scoles, P., Brown, M., Swanson, D., Holtzman, K., & Butler, A. (2010). Progress testing in clinical science education: results of a pilot project between the National Board of Medical Examiners and a US Medical School. *Medical Teacher*, *32*(6), 503-508. doi: 10.3109/01421590903514655

Dickersin, K. (1990). The existence of publication bias and risk factors for its occurrence. *JAMA*, *263*(10), 1385-1389. doi: 10.1001/jama.1990.03440100097014

Dijksterhuis, M., Scheele, F., Schuwirth, L., Essed, G., Nijhuis, J., & Braat, D. (2009). Progress testing in postgraduate medical education. *Medical Teacher*, *31*(10), e464-e468. doi: 10.3109/01421590902849545

Downing, S. (2002). Threats to the validity of locally developed multiple-choice tests in medical education: construct-irrelevant variance and construct under-representation. *Advances in Health Sciences Education*, *7*(3), 235-241. doi: https://doi.org/10.1023/A:1021112514626

Downing, S. (2004). Reliability: on the reproducibility of assessment data. *Medical Education*, *38*, 1006-1012. doi: 10.1046/j.1365-2929.2004.01932.x

Downing, S. (2005). The effects of violating standard item writing principles on tests and students: the consequences of using flawed test items on achievement examinations in medical education. *Advances in Health Sciences Education*, *10*(2), 133-143. doi: https://doi.org/10.1007/s10459-004-4019-5

Downing, S., & Haladyna, T. (2004). Validity threats: overcoming interference with proposed interpretations of assessment data. *Medical Education*, *38*(3), 327-333. doi: https://doi.org/10.1046/j.1365-2923.2004.01777.x

Drasgow, F., Levine, M., & McLaughlin, M. (1987). Detecting inappropriate test scores

with optimal and practical appropriateness indices. *Applied Psychological Measurement*, *11*, 59-79. doi: https://doi.org/10.1177/014662168701100105

Dunn, K., & Mulvenon, W. (2009). A critical review of research on formative assessment: the limited scientific evidence of the impact of formative assessment in education. *Practical Assessment, Research & Evaluation*, *14*(7), 11p.

Ecclestone, K. (2010). *Transforming formative assessment in lifelong learning*. Berkshire, UK: McGraw Hill Open University Press.

Eklöf, H. (2010). Skill and will: test-taking motivation and assessment quality. *Assessment in Education: Principles, Policy & Practice*, *17*(4), 345-56. doi: https://doi.org/10.1080/0969594X.2010.516569

Eklöf, H., & Knekta, E. (2017). Using large-scale educational data to test motivation theories: a synthesis of findings from Swedish studies on test-taking motivation. *International Journal of Quantitative Research in Education*, *4*(1/2), 52-71. doi: https://doi.org/10.1504/IJQRE.2017.086499

Finch, P. (1999). The effect of problem-based learning on the academic performance of students studying podiatric medicine in Ontario. *Medical Education*, *33*(6), 411-417. doi: https://doi.org/10.1046/j.1365-2923.1999.00347.x

Findyartini, A., Werdhani, R. A., Iryani, D., Rini, E. A., Kusumawati, R., Poncorini, E., & Primaningtyas, W. (2015). Collaborative progress test (cPT) in three medical schools in Indonesia: the validity, reliability and its use as a curriculum evaluation tool. *Medical Teacher*, *37*(4), 366-373. doi: https://doi.org/10.3109/0142159X.2014.948831

Finn, B. (2015). Measuring motivation in low-stakes assessments. *ETS Research Report Series*, 1-19. doi: https://doi.org/10.1002/ets2.12067

Finucane, P., Flannery, D., Keane, D., & Norman, G. (2010). Cross-insitutional progress testing: feasibility and value to a new medical school. *Medical Education*, *44*(2), 184-186. doi: https://doi.org/10.1111/j.1365-2923.2009.03567.x

Freeman, A., & Ricketts, C. (2010). Choosing and designing knowledge assessments: Experience at a new medical school. *Medical Teacher*, *32:7, 578-581*. doi:

10.3109/01421591003614858

Freeman, A., van der Vleuten, C., Nouns, Z., & Ricketts, C. (2010). Progress testing internationally. *Medical Teacher*, *32*(6), 451-455. doi: https://doi.org/10.3109/0142159X.2010.485231

Fullan, M. (2009). Large-scale reform comes of age. *Journal of Educational Change*, *10*(2-3), 101-112. doi: 10.1007/s10833-009-9108-z

Gesellschaft für Medizinische Ausbildung, GMA-Ausschuss Prüfungen and Kompetenzzentrum Prüfungen, Baden-Württemberg, & Fischer, M. (2008). Leitlinie für Fakultäts-interne Leistungsnachweise während des Medizinstudiums: Ein Positionspapier des GMA-Ausschusses Prüfungen und des Kompetenzzentrums Prüfungen Baden-Württemberg. *GMS Zeitschrift für Medizinische Ausbildung*, *25*(1), Doc74. Retrieved from `http://www.egms.de/de/journals/zma/2008-25/zma000558.shtml`

Given, K., Hannigan, A., & McGrath, D. (2016). Red, yellow and green: What does it mean? How the progress test informs and supports student progress. *Medical Teacher*, *38*(10), 1025-1032. doi: https://doi.org/10.3109/0142159X.2016.1147533

Haladyna, T., & Downing, S. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement*, *23*(1), 17-27. doi: https://doi.org/10.1111/j.1745-3992.2004.tb00149.x

Haladyna, T., Downing, S., & Rodriguez, M. (2010). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, *15*(2), 309-333. doi: https://doi.org/10.1207/S15324818AME1503_5

Hanß, F. (2013). *Der Leistungscheck. Entwicklung und Einführung eines fachspezifischen Progress für Zahnmediziner im Fach Anatomie* (Doctoral Thesis). Medical Faculty, University of Ulm.

Harrison, C., Könings, K., Dannefer, E., Schuwirth, L., Wass, V., & van der Vleuten, C. (2016). Factors influencing students' receptivity to formative feedback emerging from different assessment cultures. *Perspectives on Medical Education*, *5*(5), 276–284. doi: https://doi.org/10.1007/s40037-016-0297-x

Harrison, C., & Wass, V. (2016). The challenge of changing to an assessment for learning culture. *Medical Education*, *50*, 702-708. doi: https://doi.org/10.1111/medu.13058

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81-112. doi: https://doi.org/10.3102/003465430298487

Hawthorne, K., Bol, L., Pribesh, S., & Suh, Y. (2015). Effects of motivational prompts on motivation, effort, and performance on a low-stakes standardized test. *Research and Practice in Assessment*, *10*, 30-38.

Haynie III, W. (1994). Effects of multiple-choice and short-answer tests on delayed retention learning. *Journal of Technology Education*, *6*(1), 32-44. doi: 10.4066/AMJ.2011.727

Heeneman, S., Schut, S., Donkers, J., van der Vleuten, C., & Muijtjens, A. (2017). Embedding of the progress test in an assessment program designed according to the principles of programmatic assessment. *Medical Teacher*, *39*(1), 44-52. doi: 10.1080/0142159X.2016.1230183

Hochlehnert, A., Brass, K., Moeltner, A., & Jünger, J. (2011). Does medical students' preference of test format (computer-based vs. paper-based) have an influence on performance? *BMC Medical Education*, *11*, 89. doi: 10.1186/1472-6920-11-89

Hosch, B. (2012). Time on test, student motivation, and performance on the collegiate learning assessment: implications for institutional accountability. *Journal of Assessment and Institutional Effectiveness*, *2*(1), 55-76.

Irons, A. (2008). *Enhancing learning through formative assessment and feedback.* New York, NY: Routledge.

Johnson, T., Khalil, M., Peppler, R., Davey, D., & Kibble, J. (2014). Use of the NBME Comprehensive Basic Science Examination as a progress test in the preclerkship curriculum of a new medical school. *Advances in Physiology Education*, *38*(4), 315-320. doi: 10.1152/advan.00047.2014

Jünger, J., & Just, I. (2014). Recommendations of the German Society for Medical Education and the German Association of Medical Faculties regarding university-specific assessments during the study of human, dental and veterinary medicine. *GMS*

Zeitschrift für Medizinische Ausbildung, *31*(3), Doc34. doi: 10.3205/zma000926

Karay, Y., & Schauber, S. (2018). 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. *Medical Teacher*, *40*(11), 1123-1129. doi: 10.1080/0142159X.2018.1472370

Karpicke, J., & Roediger III, H. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language*, *57*(2), 151-162. doi: https://doi.org/10.1016/j.jml.2006.09.004

Kerfoot, B., Shaffer, K., McMahon, G., Baker, H., Kirdar, J., Kanter, S., ... Armstrong, E. (2011). Online spaced education progress-testing of students to confront two upcoming challenges to medical schools. *Academic Medicine*, *86*(3), 300-306. doi: 10.1097/ACM.0b013e3182087bef

Ketonen, E., & Hotulainen, R. (2019). Development of low-stakes mathematics and literacy test scores during lower secondary school – a multilevel pattern-centered analysis of student and classroom differences. *Contemporary Educational Psychology*, *59*. doi: https://doi.org/10.1016/j.cedpsych.2019.101793

Kindermann, T. (2007). Effects of naturally existing peer groups on changes in academic engagement in a cohort of sixth graders. *Child Development*, *78*(4), 1186-1203. doi: 10.1111/j.1467-8624.2007.01060.x

Kirnbauer, B., Avian, A., Jakse, N., Rugani, P., Ithaler, D., & Egger, R. (2018). First reported implementation of a German-language progress test in an undergraduate dental curriculum: A prospective study. *European Journal of Dental Education*, *22*(4), e698-e705. doi: 10.1111/eje.12381

Kluger, A., & DeNisi, A. (1996). The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*(2), 254-284. doi: 10.1111/j.1365-2923.2008.03124.x

Kostons, D., & van der Werf, G. (2015). The effects of activating prior topic and metacognitive knowledge on text comprehension scores. *British Journal of Educational Psychology*, *85*(3), 264-275. doi: 10.1111/bjep.12069

Kulik, J., & Kulik, C. (1988). Timing of feedback and verbal learning. *Review of Educational Research*, *58*(1), 79–97.

Lambert, E. (2001). College students' knowledge of human papillomavirus and effectiveness of a brief educational intervention. *Journal of the American Board of Family Practice*, *14*(3), 178-183.

Larsen, D., Butler, A., & Roediger III, H. (2008). Test-enhanced learning in medical education. *Medical Education*, *42*(10), 959-966.

Larsen, D., Butler, A., & Roediger III, H. (2013). Comparative effects of test-enhanced learning and self-explanation on long-term retention. *Medical Education*, *47*(7), 674-682. doi: 10.1111/medu.12141

Leung, S., Mok, E., & Wong, D. (2008). The impact of assessment methods on the learning of nursing students. *Nurse Education Today*, *28*(6), 711-719. doi: 10.1016/j.nedt.2007.11.004

Levine, M., & Rubin, D. (1979). Measuring the appropriateness of multiple choice test scores. *Journal of Educational Statistics*, *4*, 269-290.

Lillis, S., Yielder, J., Mogol, V., O'Connor, B., Bacal, K., Booth, R., & Bagg, W. (2014). Progress testing for medical students at the University of Auckland: Results from the first year of assessments. *Journal of Medical Education and Curricular Development*, *1*, 41-45. doi: https://doi.org/10.4137/JMECD.S20094

Liu, O., Bridgeman, B., & Adler, R. (2012). Measuring learning outcomes in higher education: Motivation matters. *Educational Researcher*, *41*(9), 352-362. doi: 10.3102/0013189X12459679

Liu, O. L., Rios, J. A., & Borden, V. (2015). The effect of motivational instruction on college students' performance on low-stakes assessment. *Educational Assessment*, *20*(2), 79-94. doi: https://doi.org/10.1080/10627197.2015.1028618

Löffler, S., Feja, C., Widmann, J., Claus, I., von Lindeman, K., & Eisnach, K. (2011). Interactive versus reproductive learning, a comparison of medical school graduates with participants of a postgraduate CPD session. *Zeitschrift für Medizinische Ausbildung*, *28*(4), Doc57. doi: 10.3205/zma000769

Löwe, B., Hartmann, M., Wild, B., Nikendei, C., Kroenke, K., Niehoff, D., ... Herzog, W. (2008). Effectiveness of a 1-year resident training program in clinical research: a controlled before-and-after study. *Journal of General Internal Medicine*, *23*(2), 122-128. doi: 10.1007/s11606-007-0397-8

Mardiastuti, H., & Werhani, R. (2011). Grade point average, progress test, and try out's test as tools for curriculum evaluation and graduates' performance prediction at the national board examination. *Journal of Medicine and Medical Sciences*, *2*(12).

Martinez, M., & Lipson, J. (1989). Assessment for learning. *Educational Leadership*, 73-75.

McDaniel, M., Anderson, J., Derbish, M., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology*, *19*(4/5), 494-513. doi: https://doi.org/10.1080/09541440701326154

McHarg, J., Bradley, P., Chamberlain, S., Ricketts, C., Searle, J., & McLachlan, J. (2005). Assessment of progress tests. *Medical Education*, *39*(2), 221-227. doi: 10.1111/j.1365-2929.2004.02060.x

Meijer, R. (1996). Person-fit research: An introduction. *Applied Measurement in Education*, *9*(1), 3-8. doi: http://dx.doi.org/10.1207/s15324818ame0901_2

Meijer, R. (2003). Diagnosing item score patterns on a test using item response theory-based person-fit statistics. *Psychological Methods*, *8*(1), 72-87.

Meijer, R., Muijtjens, A., & van der Vleuten, C. (1996). Nonparametric person-fit research: Some theoretical issues an empirical example. *Applied Measurement in Education*, *9*(1), 77-89.

Meijer, R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, *25*(2), 107-135. doi: https://doi.org/10.1177/01466210122031957

Muijtjens, A., Schuwirth, L., Cohen-Schotanus, J., Thoben, A., & van der Vleuten, C. (2008). Benchmarking by cross-institutional comparison of student achievement in a progress test. *Medical Education*, *42*(1), 82-88. doi: 10.1111/j.1365-2923.2007.02896.x

Muijtjens, A., Schuwirth, L., Cohen-Schotanus, J., & van der Vleuten, C. (2007). Origin bias of test items compromises the validity and fairness of curriculum comparisons. *Medical Education*, *41*(12), 1217-1223. doi: 10.1111/j.1365-2923.2007.02934.x

Muijtjens, A., Timmermans, I., Donkers, J., Peperkamp, R., Medema, H., Cohen-Schotanus, J., ... van der Vleuten, C. (2010). Flexible electronic feedback using the virtues of progress testing. *Medical Teacher*, *32*(6), 491-495. doi: 10.3109/0142159X.2010.486058

Neeley, S., Ulman, C., Sydelko, B., & Borges, N. (2016). The value of progress testing in undergraduate medical education: a systematic review of the literature. *Medical Science Educator*, *26*(4), 617-622. doi: https://doi.org/10.1007/s40670-016-0313-0

Nering, M., & Meijer, R. (1998). A comparison of the person response function and the lz person-fit statistic. *Applied Psychological Measurement*, *22*(1), 53-69. doi: https://doi.org/10.1177/01466216980221004

Nicol, D., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: a model and seven principles of good feedback practice. *Studies in Higher Education*, *31*(2), 199-218. doi: https://doi.org/10.1080/03075070600572090

Norcini, J., Anderson, M., Bollela, V., Burch, V., Costa, M., Duvivier, R., ... Swanson, D. (2018). 2018 consensus framework for good assessment. *Medical Teacher*, *40*(11), 1102-1109. doi: 10.1080/0142159X.2018.1500016

Norman, G., Neville, A., Blake, J., & Mueller, B. (2010). Assessment steers learning down the right road: impact of progress testing on licensing examination performance. *Medical Teacher*, *32*(6), 496-499. doi: https://doi.org/10.3109/0142159X.2010.486063

Nouns, Z., & Brauns, K. (2008). Prüfungen auf die Agenda - Hochschuldidaktische Perspektiven auf Reformen im Prüfungswesen. In S. Dany, B. Szczyrba, & J. Wildt (Eds.), (p. 114-128). Bielefeld: W. Bertelsmann Verlag.

Nouns, Z., & Georg, W. (2010). Progress testing in German speaking countries. *Medical Teacher*, *32*(6), 467-470. doi: https://doi.org/10.3109/0142159X.2010.485656

Nouns, Z., Hanfler, S., Brauns, K., Föller, T., Fuhrmann, S., Kölbel, A., Mertens, ...

Osterberg, K. (2004). Do progress tests predict the outcome of national exams? (short communication: 2f 3). In *AMEE-conference, 5-8 September 2004, Edinburgh, UK.*

OECD. (1999). *Measuring student knowledge and skills. A new framework for assessment* (Tech. Rep.). Organisation for Economic Co-Operation and Development. Retrieved from `https://www.oecd.org/edu/school/programmeforinternationalstudentassessmentpisa/33693997.pdf`

OECD. (2003). *The PISA 2003 assessment framework. Mathematics, reading, science and problem solving knowledge and skills* (Tech. Rep.). Organisation for Economic Co-Operation and Development. Retrieved from `https://www.oecd.org/edu/school/programmeforinternationalstudentassessmentpisa/33694881.pdf`

O'Neil, J., H.F., Sugrue, B., & Baker, E. (1995). Effects of motivational interventions on the National Assessment of Education Progress Mathematics Performance. *Educational Assessment*, *3*(2), 135-157. doi: https://doi.org/10.1207/s15326977ea0302_2

Osterberg, K., Kölbel, S., & Brauns, K. (2006). Der Progress Test Medizin: Erfahrungen an der Charité Berlin. *GMS Zeitschrift für Medizinische Ausbildung*, *23*(3), Doc46.

Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology*, *8*, 422. doi: 10.3389/fpsyg.2017.00422

Patall, E. A., Cooper, H., & Robinson, J. C. (2008). The effects of choice on intrinsic motivation and related outcomes: A meta-analysis of research findings. *Psychological Bulletin*, *134*(2), 270-300. doi: https ://doi.org/10.1037/0033-2909.134.2.270

Peeraer, G., De Winter, B., Muijtjens, A., Remmen, R., Bossaert, L., & Scherpbier, A. (2009). Evaluating the effectiveness of curriculum change. Is there a difference between graduating student outcomes from two ifferent curricula? *Medical Teacher*, *31*(3), e64-e68. doi: 10.1080/01421590802512920

Penk, D., C. amd Richter. (2017). Change in test-taking motivation and its relationship to test performance in low-stakes assessments. *Educational Assessment, Evaluation and Accountability*, *29*(1). doi: http://dx.doi.org/10.1007/s11092-016-9249-6

Plessas, A. (2015). Validity of progress testing in healthcare education. *International Journal of Humanities Social Sciences and Education*, *2*(8), 23-33.

Portanova, R., Adelman, M., Jollick, J., Schuler, S., Modrzakowski, M., Soper, E., & Ross-Lee, B. (2000). Student assessment in the Ohio University College of Osteopathic Medicine CORE system: progress testing and objective structured clinical examinations. *Journal of American Osteopathy Association*, *100*(11), 707-712.

Prince, K., van Mameren, H., Hylkema, N., Drukker, J., Scherpbier, A., & van der Vleuten, C. (2003). Does problem-based learning lead to deficiencies in basic science knowledge? An empirical case on anatomy. *Medical Education*, *37*(1), 15-21. doi: https://doi.org/10.1046/j.1365-2923.2003.01402.x

Pugh, D., & Regehr, G. (2016). Taking the sting out of assessment: is there a role for progress testing? *Medical Education*, *50*(7), 721-729. doi: 10.1111/medu.12985

Pugh, D., Touchie, C., Wood, T., & Humphrey-Murto, S. (2014). Progress testing: is there a role for the osce? *Medical Education*, *48*(6), 623-631. doi: 10.1111/medu.12423

Rademakers, J., Ten Cate, T., & Bär, P. (2005). Progress testing with short answer questions. *Medical Teacher*, *27*(7), 578-582. doi: 10.1080/01421590500062749

Raman, M., McLaughlin, K., Violato, C., Rostom, A., Allard, J., & Coderre, S. (2010). Teaching in small portions dispersed over time enhances long-term knowledge retention. *Medical Teacher*, *32*(3), 250-255. doi: 10.3109/01421590903197019

Ramraje, S., & Sable, P. (2011). Comparison of the effect of post-instruction multiple-choice and short-answer tests on delayed retention learning. *Australasian Medical Journal*, *4*(6), 332-339. doi: 10.4066/AMJ.2011.727

Raupach, T., Grefe, C., Brown, J., Meyer, K., Schuelper, N., & Anders, S. (2015). Moving knowledge acquisition from the lecture hall to the student home: a prospective intervention study. *Journal of Medical Internet Research*, *17*(9), e223. doi: 10.2196/jmir.3814

Ravesloot, C., van der Schaaf, M., Haaring, C., Kruitwagen, C., Beek, E., Ten Cate, O., & van Schaik, J. (2012). Construct validation of progress testing to measure knowledge and visual skills in radiology. *Medical Teacher*, *34*(12), 1047-1055. doi:

10.3109/0142159X.2012.716177

Renkl, A., Stark, R., Gruber, H., & Mandl, H. (1998). Learning from worked-out examples: the effects of example variability and elicited self-explanations. *Contemporary Educational Psychology*, *23*(1), 90-108. doi: https://doi.org/10.1006/ceps.1997.0959

Reynolds, L., & Kostich, S. (2017). Taking the sting out of assessment: is there a role for progress testing? *Medical Education*, *51*(7), 768. doi: 10.1111/medu.13259

Rheinberg, F., Vollmeyer, R., & Burns, B. (2001). FAM: Ein Fragebogen zur Erfassung aktueller Motivation in Lern- und Leistungssituationen [a questionnaire to assess current motivation in learning situations]. *Diagnostica*, *47*(2), 57-66. doi: https://doi.org/10.1026//0012-1924.47.2.57

Ricketts, C., Freeman, A., Pagliuca, G., Coombes, L., & Archer, J. (2010). Difficult decisions for progress testing: How much and how often? *Medical Teacher*, *32*(6), 513-515. doi: 10.3109/0142159X.2010.485651

Ringsted, C., Hodges, B., & Scherpbier, A. (2011). 'The research compass': An introduction to research in medical education: AMEE Guide No. 56. *Medical Teacher*, *33*(9), 695-709. doi: 10.3109/0142159X.2011.595436

Roediger III, H., & Butler, A. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*(1), 20-27. doi: 10.1016/j.tics.2010.09.003

Rushton, A. (2005). Formative assessment: a key to deep learning? *Medical Teacher*, *27*(6), 509-513. doi: 10.1080/01421590500129159

Rutgers, D., van Raamt, F., van Lankeren, W., Ravesloot, C., van der Gijp, A., Ten Cate, T., & van Schaik, J. (2018). Fourteen years of progress testing in radiology residency training: experiences from The Netherlands. *European Radiology*, *28*(5), 2208-2215. doi: 10.1007/s00330-017-5138-8

Rutkowski, D., & Wild, J. (2015). Stakes matter: student motivation and the validity of student assessments for teacher evaluation. *Educational Assessment*, *20*(3), 165-179. doi: https://doi.org/10.1080/10627197.2015.1059273

Ruzek, E., Hafen, C., Allen, J., Gregory, A., Mikami, A., & Pianta, R. (2016). How

teacher emotional support motivates students: The mediating roles of perceived peer relatedness, autonomy support, and competence. *Learning and Instruction*, *42*, 95-103. doi: 10.1016/j.learninstruc.2016.01.004

Ryan, R., & Deci, E. (2000). Intrinsic and extrinsic motivations: classic definitions and new directions. *Contemporary Educational Psychology*, *25*(1), 54-67. doi: https://doi.org/10.1006/ceps.1999.1020

Scargle, J. (2000). Publication bias: the 'file-drawer problem' in scientific inference. *Journal of Scientific Exploration*, *14*(2), 94–106.

Schaap, L., Schmidt, H., & Verkoeijen, P. (2012). Assessing knowledge growth in a psychology curriculum: which students improve most? *Assessment & Evaluation in Higher Education*, *31*(7), 875-887. doi: https://doi.org/10.1080/02602938.2011.581747

Schauber, S., & Nouns, Z. (2010). Using the cumulative deviation method for cross-institutional benchmarking in the Berlin progress test. *Medical Teacher*, *32*(6), 471-475. doi: 10.3109/0142159X.2010.485653

Schmidmaier, R., Ebersbach, R., Schiller, M., Hege, I., Holzer, M., & Fischer, M. (2011). Using electronic flashcards to promote learning in medical students: retesting versus restudying. *Medical Education*, *45*(11), 1101-1110. doi: 10.1111/j.1365-2923.2011.04043.x

Schmidmaier, R., Holzer, M., Angstwurm, M., Nouns, Z., Reincke, M., & Fischer, M. (2010). Using the Progress Test Medizin (PTM) for evaluation of the Medical Curriculum Munich (MeCuM). *GMS Zeitschrift für Medizinische Ausbildung*, *27*(5), Doc70. doi: 10.3205/zma000707

Schmitt, N., Chan, D., Sacco, J., McFarland, L., & Jennings, D. (1999). Correlates of person fit and effect of person fit on test validity. *Applied Psychological Measurement*, *23*(1), 41-53. doi: https://doi.org/10.1177/01466219922031176

Schut, S., Driessen, E., van Tartwijk, J., van der Vleuten, C., & Heeneman, S. (2018). Stakes in the eye of the beholder: an international study of learners' perceptions within programmatic assessment. *Medical Education*, *52*(6), 654-663. doi:

10.1111/medu.13532

Schüttpelz-Brauns, K. (2017). Umgang mit nichtseriösen Teilnehmern bei nicht bestehensrelevanten Tests. *Empirische Evaluationsmethoden (Wissenschaftliche Veranstaltungen)*, *7*, 83-97.

Schuwirth, L., Bosman, G., Henning, R., Rinkel, R., & Wenink, A. (2010). Collaboration on progress testing in medical schools in the Netherlands. *Medical Teacher*, *32*(6), 476-479. doi: 10.3109/0142159X.2010.485658

Schuwirth, L., & van der Vleuten, C. (2012). The use of progress testing. *Perspectives in Medical Education*, *1*(1), 24-30. doi: 10.1007/s40037-012-0007-2

Scouller, K. (1998). The influence of assessment method on students' learning approaches: Multiple choice question examination versus assignment essay. *Higher Education*, *35*(4), 453-472. doi: https://doi.org/10.1023/A:1003196224280

Setzer, J. C., Wise, S. L., van den Heuvel, J. R., & Ling, G. (2013). An investigation of examinee test-taking effort on a large-scale assessment. *Applied Measurement in Education*, *26*(1), 34-49. doi: https://doi.org/10.1080/08957347.2013.739453

Seybert, A., & Barton, C. (2007). Simulation-based learning to teach blood pressure assessment to doctor of pharmacy students. *American Journal of Pharmaceutical Education*, *71*(3), 48.

Shute, V. (2008). Focus on formative feedback. *Review of Educational Research*, *78*(1), 153-189. doi: 10.3102/0034654307313795

Siegling-Vlitakis, C., Stephan Birk, S., Kröger, A., Matenaers, C., Beitz-Radzio, C., Staszyk, C., ... Ehlers, J. (2014). PTT: Progress Test Tiermedizin - Ein individuelles Feedback-Werkzeug für Studierende. *Deutsches Tierärzteblatt*, *08*, 1076-1082.

Skinner, B. (1958). Teaching machines. *Science*, *128*(3330), 969–977.

Smyth, K. (2004). The benefits of students learning about critical evaluation rather than being summatively judged. *Assessment & Evaluation in Higher Education*, *29*(3), 369-377. doi: https://doi.org/10.1080/0260293042000197609

Soliman, M., Al-Shaikh, G., & Alnassar, S. (2016). Use of cross-institutional progress test as a predictor of performance in a new medical college. *Advances in Medical*

Education and Practice, *7*. doi: https://doi.org/10.2147/AMEP.S89643

Streiner, D., Norman, G., & Cariney, J. (2015). *Health measurement scales - a practical guide to their development and use* (5th ed.). Oxford: University Press.

Sundre, D., & Kitsantas, A. (2004). An exploration of the psychology of the examinee: Can examinee self-regulation and test-taking motivation predict consequential and non-consequential test performance? *Contemporary Educational Psychology*, *29*(1), 6-26. doi: https://doi.org/10.1016/S0361-476X(02)00063-2

Sundre, D. L., & Moore, D. L. (2002). The student opinion scale: A measure of examinee motivation. *Assessment Update*, *14*(1), 8–9.

Swanson, D., Holtzman, K., Butler, A., Langer, M., Nelson, M., Chow, J., ... The Multi-School Progress Testing Committee (2010). Collaboration across the pond: The multi-school progress testing project. *Medical Teacher*, *32*(6), 480-485. doi: 10.3109/0142159X.2010.485655

Swerdzewski, P., Harmes, J., & Finney, S. (2011). Two approaches for identifying low-motivated students in a low-stakes assessment context. *Applied Measurement in Education*, *24*(2), 162-188. doi: https://doi.org/10.1080/08957347.2011.555217

Thelk, A., Sundre, D., Horst, S., & Finney, S. (2009). Motivation matters: using the student opinion scale to make valid inferences about student performance. *The Journal of General Education*, *58*(3), 129-151. doi: https://doi.org/10.1353/jge.0.0047

Tio, R., Schutte, B., Meiboom, A., Greidanus, J., Dubois, E., Bremers, A., & Dutch Working Group of the Interuniversity Progress Test of Medicine. (2016). The progress test of medicine: the Dutch experience. *Perspectives in Medical Education*, *5*(1), 51-55. doi: 10.1007/s40037-015-0237-1

Tomic, E., Martins, M., Lotufo, P., & Benseñor, I. (2005). Progress testing: evaluation of four years of application in the school of medicine, University of Sã o Paulo. *Clinics (Sao Paulo)*, *60*(5), 389-396. doi: http://dx.doi.org/10.1590/S1807-59322005000500007

Tuten, T., Galesic, M., & Bosnjak, M. (2004). Effects of immediate versus delayed notification of prize draw results and announced survey duration on response behavior

in web surveys - an experiment. *Social Science Computer Review*, *22*(3), 377-384. doi: 10.1177/0894439304265640

van der Vleuten, C. (1996). The assessment of professional competence: Developments, research and practical implications. *Advances in Health Sciences Education*, *1*, 41-67. doi: 10.1007/BF00596229

van der Vleuten, C., Freeman, A., & Collares, C. (2018). Progress test utopia. *Perspectives on Medical Education*, *7*(2), 136-138. doi: 10.1007/s40037-018-0413-1

van der Vleuten, C., Scherpbier, A., Dolmans, D., Schuwirth, L., Verwijnen, G., & Wolfhagen, H. (2000). Clerkship assessment assessed. *Medical Teacher*, *22*(6), 592-600. doi: 10.1080/01421590050175587

van der Vleuten, C., Schuwirth, L., Driessen, E., Govaerts, M., & Heeneman, S. (2015). Twelve tips for programmatic assessment. *Medical Teacher*, *37*(7), 641-646. doi: 10.3109/0142159X.2014.973388

van der Vleuten, C., Schuwirth, L., Muijtjens, A., Thoben, A., Cohen-Schotanus, J., & van Boven, C. (2004). Cross institutional collaboration in assessment: a case on progress testing. *Medical Teacher*, *26*(8), 719-725. doi: 10.1080/01421590400016464

van der Vleuten, C., Schuwirth, L., Scheele, F., Driessen, E., & Hodges, B. (2010). The assessment of professional competence: building blocks for theory development. *Clinical Obstetrics and Gynecology*, *24*(6), 703-719. doi: 10.1016/j.bpobgyn.2010.04.001

van der Vleuten, C., Verwijnen, G., & Wijnen, W. (1996). Fifteen years of experience with progress testing in a problem-based learning curriculum. *Medical Teacher*, *18*(2), 103-109. doi: https://doi.org/10.3109/01421599609034142

van Berkel, H. (1990). Assessment in a problem-based medical curriculum. *Higher Education*, *19*(2), 123-146. doi: https://doi.org/10.1007/BF00137104

van Berkel, H., Nuy, H., & Geerligs, T. (1995). The influence of progress tests and block tests on study behaviour. *Instructional Science*, *22*(4), 317-333. doi: https://doi.org/10.1007/BF00891784

van der Veken, J., Valcke, M., De Maeseneer, J., Schuwirth, L., & Derese, A. (2009). Impact on knowledge acquisition of the transition from a conventional to an in-

tegrated contextual medical curriculum. *Medical Education*, *43*(7), 704-713. doi: 10.1111/j.1365-2923.2009.03397.x

van Diest, R., van Dalen, J., Bak, M., Schruers, K., van der Vleuten, C., Muijtjens, A., & Scherpbier, A. (2004). Growth of knowledge in psychiatry and behavioural sciences in a problem-based learning curriculum. *Medical Education*, *38*(12), 1295-1301. doi: 10.1111/j.1365-2929.2004.02022.x

Verhoeven, B. (2003). *Progress testing - the utility of an assessment concept* (Dissertation). Universiteit Maastricht.

Verhoeven, B., Snellen-Balendong, H., Hay, I., Boon, J., van der Linde, M., Blitz-Lindeque, J., ... van der Vleuten, C. (2005). The versatility of progress testing assessed in an international context: a start for benchmarking global standardization? *Medical Teacher*, *27*(6), 514-520. doi: https://doi.org/10.1080/01421590500136238

Verhoeven, B., Verwijnen, G., Scherpbier, A., Holdrinet, R., Oeseburg, B., Bultec, J., & van der Vleuten, C. (1998). An analysis of progress test results of pbl and non-pbl students. *Medical Teacher*, *20*(4), 310-331. doi: https://doi.org/10.1080/01421599880724

Verhoeven, B., Verwijnen, G., Scherpbier, A., Schuwirth, L., & van der Vleuten, C. (1999). Quality assurance in test construction: the approach of a multidisciplinary central test committee. *Education for Health*, *12*(1), 49-60.

Verhoeven, B., Verwijnen, G., Scherpbier, A., & van der Vleuten, C. (2002). Growth of medical knowledge. *Medical Education*, *36*(8), 711-717. doi: 10.1046/j.1365-2923.2002.01268.x

Wade, L., Harrison, C., Hollands, J., Mattick, K., Ricketts, C., & Wass, V. (2012). Student perceptions of the progress test in two settings and the implications for test development. *Advances in Health Sciences Education*, *17*, 573-583. doi: 10.1007/s10459-011-9334-z

Wallach, P., Crespo, L., Holtzman, K., Galbraith, R., & Swanson, D. (2006). Use of a committee review process to improve the quality of course examinations. *Advances in Health Sciences Education*, *11*(1), 61-68. doi: 10.1007/s10459-004-7515-8

Warburton, V. (2017). Peer and teacher influences on the motivational climate in physical education: A longitudinal perspective on achievement goal adoption. *Contemporary Educational Psychology*, *51*(10), 303-314. doi: https://doi.org/10.1016/j.cedpsych.2017.08.001

Ware, J., & Vik, T. (2009). Quality assurance of item writing: during the introduction of multiple choice questions in medicine for high stakes examinations. *Medical Teacher*, *31*(3), 238-243. doi: 10.1080/01421590802155597

Waskiewicz, R. (2011). Pharmacy students' test-taking motivation-effort on a low-stakes standardized test. *American Journal of Pharmaceutical Education*, *75*(3), 1-8.

Wentzel, K., Battle, A., Russell, S., & Looney, L. (2010). Social supports from teachers and peers as predictors of academic and social motivation. *Contemporary Educational Psychology*, *35*(3), 193-202. doi: https://doi.org/10.1016/j.cedpsych.2010.03.002

Wentzel, K., Muenks, K., McNeisha, D., & Russell, S. (2017). Peer and teacher supports in relation to motivation and effort: A multi-level study. *Contemporary Educational Psychology*, *49*, 32-45. doi: https://doi.org/10.1016/j.cedpsych.2016.11.002

Wetzels, S., Kester, L., van Merriënboer, J., & Broers, N. (2011). The influence of prior knowledge on the retrieval-directed function of note taking in prior knowledge activation. *British Journal of Educational Psychology*, *81*(2), 274-291. doi: 10.1348/000709910X517425

White, C., & Gruppen, L. (2010). Understanding medical education: Evidence, theory and practice. In T. Swanwick (Ed.), (p. 271-282). Chichester (West Sussex): Wiley-Blackwell.

Wigfield, A., & Eccles, J. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology*, *25*(1), 68-81. doi: https://doi.org/10.1006/ceps.1999.1015

Willoughby, T., Dimond, E., & Smull, N. (1977). Correlation of quarterly profile examination and National Board of Medical Examiner scores. *Educational and Psychological Measurement*, *37*(2), 445-449. doi: https://doi.org/10.1177/001316447703700219

Willoughby, T., & Hutcheson, S. (1978). Edumetric validity of the quarterly profile examination. *Educational and Psychological Measurement*, *38*(4), 1057-1061. doi: https://doi.org/10.1177/001316447803800425

Wise, S. (2009). Strategies for managing the problem of unmotivated examinees in low-stakes testing programs. *Journal of General Education*, *58*(3), 152-166.

Wise, S., & DeMars, C. (2005). Low examinee effort in low-stakes assessment: problems and potential solutions. *Educational Assessment*, *10*(1), 1-17. doi: https://doi.org/10.1207/s15326977ea1001_1

Wise, S., & DeMars, C. (2010). Examinee non-effort and the validity of program assessment results. *Educational Assessment*, *15*(1), 27-41. doi: https://doi.org/10.1080/10627191003673216

Wise, S., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*(2), 163-183. doi: https://doi.org/10.1207/s15324818ame1802_2

Wise, S. L., & Ma, L. (2012). Setting response time thresholds for a CAT item pool: The normative threshold method. In *Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, Canada.*

Wolf, L. F., & Smith, J. K. (1995). The consequence of consequence: Motivation, anxiety, and test performance. *Applied Measurements in Education*, *8*(3), 227-242. doi: https://doi.org/10.1207/s15324818ame0803_3

Wood, D. (2010). Understanding medical education: Evidence, theory and practice. In T. Swanwick (Ed.), (p. 259-270). Chichester (West Sussex): Wiley-Blackwell.

Wrigley, W., van der Vleuten, C., Freeman, A., & Muijtjens, A. (2012). A systematic framework for the progress test: Strengths, constraints and issues: AMEE guide no. 71. *Medical Teacher*, *34*(9), 683-697. doi: 10.3109/0142159X.2012.704437

Yielder, J., Wearn, A., Chen, Y., Henning, M., Weller, J., Lillis, S., ... Bagg, W. (2017). A qualitative exploration of student perceptions of the impact of progress tests on learning and emotional wellbeing. *BMC Medical Education*, *17*(1), 148. doi: 10.1186/s12909-017-0984-2

Zheng, M., & Bender, D. (2018). Evaluating outcomes of computer-based classroom testing: Student acceptance and impact on learning and exam performance. *Medical Teacher*, *41*(1), 75-82. doi: 10.1080/0142159X.2018.1441984

Acknowledgments

Since I started working on the Berlin Progress Test in 2003, I have been dealing with low-stakes assessments. I am particularly fascinated by the tension between existing curricula, which are strongly focused on grades, and the increasing demand for self-regulated learning with individual learning paths, including formative low-stakes assessments. In my doctoral thesis I focused on the statistical identification of non-serious test-takers to provide serious participants with more valid feedback on their level of knowledge. In recent years, my perspective has broadened to the utility of low-stakes assessments, especially conditions under which self-regulated learning will be possible in the long term.

The present work is not an individual achievement. Rather, there were many people who accompanied my path and supported me in their special ways.

First of all, I would like to thank the people who made it possible for me to carry out and publish the studies for my habilitation despite the change of my main subject in the meantime. Prof. Dr. Dr. h.c. Dr. h.c. Herta Flor supervised my habilitation and had an open ear. She gave me direction whenever I was lost. Prof. Dr. emeritus Bodo Krause taught me research methodologies and follow a systematic approach. He gave me roots. Prof. Dr. Udo Obertacke and Dr. Harald Fritz-Joas gave me the freedom to research, to develop myself and supported me whenever I needed help. They gave me wings.

Without the members of the BPT team and the cooperative partners of the Berlin Progress Test it would not have been possible to carry out or complete the studies even after the change of my job. Therefore, I would like to thank all those who were involved in the studies, be it as co-thinkers and co-authors or as providers of data.

Furthermore, I want to thank my silent helpers who are otherwise unseen. Philipp

Wirtz patiently did literature research over and over again, whenever I asked him, regardless of the strange issue I identified for him. Kathrin Nühse I want to thank for critically reading and editing the first draft of this manuscript. Anne Berwanger changed my peculiar German English into proper English.

My special thanks go to my parents. They have encouraged and supported me all my life, no matter what kind of silly ideas I had in my head. They were and are always there for me. I thank you!

Finally, but no less sincerely, I thank my family and friends for everything you are for me.