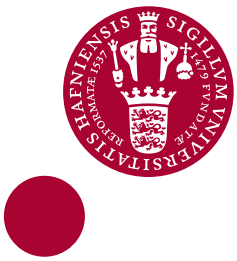

Development and validation of a multiple-choice questionnaire-based theoretical test in direct ophthalmoscopy

Jørgensen, Morten; Savran, Mona Meral; Christakopoulos, Christos; Bek, Toke; Grauslund, Jakob; Toft, Peter Bjerre; Ziemssen, Focke; Konge, Lars; Sørensen, Torben Lykke; Subhi, Yousif

Published in:
Acta Ophthalmologica

DOI:
10.1111/aos.14065

Publication date:
2019

Document version
Early version, also known as pre-print

Development and validation of a multiple-choice questionnaire-based theoretical test in direct ophthalmoscopy

Authors

Morten Jørgensen[1,2,3], Mona Meral Savran[2,4], Christos Christakopoulos[5], Toke Bek[6], Jakob Grauslund[7,8], Peter Bjerre Toft[9], Focke Ziemssen[10], Lars Konge[2,3], Torben Lykke Sørensen[1,3], Yousif Subhi[1,3]

Affiliations

1. Department of Ophthalmology, Zealand University Hospital, Roskilde, Denmark.

2. CAMES - Copenhagen Academy for Medical Education and Simulation, Capital Region of Denmark, Copenhagen, Denmark.

3. Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark.

4. Department of Obstetrics and Gynaecology, Copenhagen University Hospital Amager and Hvidovre, Hvidovre, Denmark.

5. Department of Ophthalmology, Zealand University Hospital, Næstved, Denmark.

6. Department of Ophthalmology, Aarhus University Hospital, Aarhus, Denmark.

7. Department of Ophthalmology, Odense University Hospital, Odense, Denmark.

8. Department of Clinical Research, Faculty of Healthy Science, University of Southern Denmark, Odense, Denmark.

9. Department of Ophthalmology, Rigshospitalet, Copenhagen University Hospital, Copenhagen, Denmark.

10. Center for Ophthalmology, Eberhard-Karl University Tübingen, Tübingen, Germany.

Correspondence

Yousif Subhi, MD PhD
Department of Ophthalmology
Zealand University Hospital, Roskilde
Vestermarksvej 23. DK-4000 Roskilde. Denmark
Tel: +45 47323900
Fax: +45 46362645
Email: ysubhi@gmail.com

Abstract

Purpose: Direct ophthalmoscopy can reveal systemic, neurologic, and ophthalmic conditions; but is poorly mastered among young physicians. A theoretical test is needed to measure effect of educational interventions. We developed and gathered validity evidence for a multiple-choice questionnaire (MCQ)-based theoretical test in direct ophthalmoscopy.

Methods: The MCQ was developed by interviewing experts. Then, validity evidence was evaluated using Messick's validity framework. *Content* was ensured by inviting the experts to contribute in a Delphi-like process. *Response process* was ensured by piloting and by streamlining all instructions. Then, the test was taken by ophthalmologists and by medical students without experience in direct ophthalmoscopy. Results were used to evaluate *internal structure* (item quality analysis and internal consistency), *relations to other variables* (correlation of test scores to experience level), and *consequences* (establishment of pass-fail score and the consequences of its use).

Results: The first phase of the study yielded 100 MCQs. In second phase, we identified that 60 items fulfilled predefined relevance and item quality requirements. These items demonstrated very high internal consistency (Cronbach's alpha = 0.95), significantly discriminated medical students from specialists ($P < 0.001$, independent samples t-test), and the established pass-fail score of 50 (83 %) correct answers resulted in no false positives (students passing) and no false negatives (specialists failing). A Decision study identified that sampling 15 items suffice for certification.

Conclusion: We developed and validated an MCQ-based theoretical test in direct ophthalmoscopy that enables an evidence-based approach to measuring, evaluating, and certifying the theoretical knowledge necessary for direct ophthalmoscopy.

Key Words: Direct ophthalmoscopy, multiple-choice questionnaire, theoretical test, education, Messick.

Introduction

Direct ophthalmoscopy is an important part of the medical curriculum that enables basic examination of the posterior section of the eye. Direct ophthalmoscopy can be performed quickly and reveals a number of sight- and life-threatening conditions (Bruce et al. 2011, Ting et al. 2016, Morad et al. 2004, Mackay et al. 2015), wherein rapid diagnosis is crucial for best possible clinical outcomes (Leske et al. 2003, Bacon et al. 1993, Chawla et al. 2016, Rasmussen et al. 2015, Patil et al. 2016, Georgouli et al. 2011, Eijk et al. 2016). Proficiency in the skill is expected of all young medical graduates (International Task Force on Opthalmic Education of Medical Students 2006). However, direct ophthalmoscopy is one of the poorest mastered clinical skills among young physicians (Ringsted et al. 2010). Studies have revealed that the lack of proficiency and confidence in the skill leads to avoidance of the ophthalmoscopic examination in the clinical practice (Bruce et al. 2011, Ringsted et al. 2010, Nicholl et al. 2012, Gupta & Lam 2006).

The acquisition of cognitive skills provides trainees with a theoretical foundation to perform and improve technical skills (Kohls-Gatzoulis et al. 2004). A test is needed to ensure sufficient theoretical knowledge, and testing also improves later retention of information, which is also known as the testing effect (Kromann et al. 2009). Important decisions regarding type, format, content, validity, reliability, and cost-effectiveness need to be made when developing a test (Schuwirth & van der Vleuten 2003). Written tests are more cost-effective and reliable than other assessment types. Written tests can have different formats. Multiple-choice questions (MCQs) have several advantages: MCQs can assess a large area of knowledge, are reproducible, have high reliability, and have relatively lower answering and score time (Schuwirth & van der Vleuten 2003).

To our knowledge, no such MCQ is available for testing theoretical knowledge in direct ophthalmoscopy, which have been developed using best-practices for MCQ-development and where evidence of validity has been collected using any modern validity framework. If MCQs are not carefully developed using evidence-based best-practices, they can be flawed and results obtained from such flawed MCQs may not measure what is intended to measure. Collecting evidence of validity allows one to evaluate to what extend a test measures what is intended to measure. In this study, we developed a theoretical test of proficiency in direct ophthalmoscopy using best-practice in MCQ

development and collected validity evidence using Messick's validity framework (Downing & Yudkowsky 2009, Thomsen et al. 2015).

Methods

Study design

Our study consisted of two consecutive phases. In the first phase, we developed the MCQ test. For this part of the study, we recruited five content experts from five different centers (from Denmark and Germany). Four were clinical professors in ophthalmology and one was a clinical associate professor in ophthalmology who is head of the ophthalmology course for medical students. All content experts were active instructors in ophthalmology and direct ophthalmoscopy in medical schools. In the second phase, we collected validity evidence for the developed MCQ test. For this part of the study, we recruited participants defined as either inexperienced novices (medical students with no experience in direct ophthalmology) or experienced specialists (specialists in ophthalmology). Medical students were eligible if they had clinical training (i.e. they must have a conceptual understanding of the importance of clinical examination) but not received any ophthalmoscopy training. Specialists in ophthalmology were defined as those who had completed their residency.

Our study protocol was presented to the ethics committee of the Capital Region of Denmark, which deemed that approval was unnecessary due to the nature of the study and gave a waiver (jr. no. 17028552). All aspects of the study followed the ethical principles of the Declaration of Helsinki. Participants were informed about the study and gave informed consent prior to participation.

Phase 1: Development of the test

We first reviewed the literature, including several examples of books used in medical education of ophthalmology (Harper 2016, Fahmy et al. 2013, Bek et al. 2016). In addition, we also reviewed the Principles and Guidelines of a Curriculum for Ophthalmic Education of Medical Students by the International Task Force on Ophthalmic Education of Medical Students (2006) of the International Council of Ophthalmology. This publication is a consensus statement from multiple international panels established to facilitate streamlined curricula for training of medical students in ophthalmology. This consensus statement includes knowledge and skills that are considered necessary when training medical students in ophthalmology, and it also includes the use of direct ophthalmoscopy in different scenarios of clinical practice.

We conducted unstructured interviews with content experts (n=5) based on the reviewed literature. We initiated the interviews with the question *"What do you consider as relevant theoretical knowledge when performing a direct ophthalmoscopic examination?"* and followed up with elaborative questions on the themes and issues raised by the experts. Interviews were recorded, transcribed, and analyzed to identify themes and issues, upon which questions and items were constructed. All questions and items were constructed according to the MCQ guidelines by Case and Swanson (2001), and inspired by Haladyna and Rodriguez (2013). The selected item format was one-best-answer questions with a stem consisting all necessary information and three options with one best answer and two misleading options (Case & Swanson 2001, Haladyna & Rodriguez 2013). All three options were made as homogeneous as possible and within the same thematic category.

Phase 2: Validity evidence of the test

We collected validity evidence by following the contemporary framework of validity by Samuel Messick (Downing & Yudkowsky 2009, Thomsen et al. 2015). Messick describes five sources of validity evidence: content, response process, internal structure, relationship to other variables, and consequences. Each is described briefly in the following with a specific strategy to collecting validity evidence.

Content *(i.e. relevance of the test content to different aspects of direct ophthalmoscopy)*: Four content experts evaluated the MCQs in a Delphi-like process, where all constructed MCQ items were commented and evaluated until the finally phrased MCQ item were rated on a scale from 1 (completely irrelevant) to 5 (extremely relevant) (Savran et al. 2014, Savran et al. 2015, Jensen et al. 2016). Experts were motivated to comment the phrasing, the clarity, and to suggest new MCQ items. Rephrased items were sent for re-evaluation. We defined relevance as those items receiving an average score of ≥ 3 and where no content expert rated 1 (in other words, items were discarded if rated 1 by at least one expert or rated <3 on average). These selection criteria were not revealed to the content experts to avoid bias. Included MCQ items were piloted on four students, wherein the final wording and clarity of the items were ensured.

Response process *(i.e. elimination or control of potential sources of bias)*: After the pilot study, all MCQ items were answered by medical students and by specialists in ophthalmology. Medical

students were recruited through advertising on social media. Specialists were recruited from the Department of Ophthalmology at Zealand University Hospital. The same test instructor presented the test to all participants based on pre-defined set of instructions (participants were supervised for the duration of the test, and they were not allowed to use handbooks, access web resources, or ask for help) to ensure streamlining of the response process.

Internal structure *(i.e. degree to which different items that measure comparable constructs produce consistent results)*: Power calculation was not made since it was not possible to make qualified assumptions towards scores on the newly developed test. We decided to recruit 10 specialists in ophthalmology and 20 medical students to assume normal distribution (≥10 participants in each proficiency level (Bloch & Norman 2012)) and to be able to detect clinically relevant differences (i.e. our test to be relevant it must be able to discriminate between a typical class of students and 10 specialists). After testing individuals of different competency levels (medical students and specialists in ophthalmology), we used the data to evaluate item quality based on item difficulty and item discrimination (Downing & Yudkowsky 2009). Level I items are considered to be the best fit, and they are of middle difficulty (item difficulty: 0.45 to 0.75) with high discriminatory ability (item discrimination ≥ 0.20) (Haladyna & Rodriguez 2013). Level II items are considered to be the next in line, and are relatively easy questions (item difficulty: 0.76 to 0.91) with high discriminatory ability (item discrimination ≥ 0.15) (Haladyna & Rodriguez 2013). Level III items are difficult items (item difficulty: 0.25 to 0.44) with some discriminatory ability (item discrimination ≥ 0.10), which are preferably not included in a test (Haladyna & Rodriguez 2013). Level IV items are the rest with the poorest quality (item difficulty: <0.24 or >0.91, regardless of item discrimination) (Haladyna & Rodriguez 2013). Item level III and level IV were discarded, as is recommended for MCQ development (Downing & Yudkowsky 2009). The remaining MCQ items were subject to reliability testing to evaluate consistency quality.

Relationship to other variables *(i.e. correlation of test scores to other external variables such as level of competence)*: We evaluated relationship to other variables by comparing test results between groups from different competency levels (medical students and specialists in ophthalmology).

Consequences *(i.e. consequence of obtaining a certain test score)*: We established a pass-fail standard using the contrasting groups' method on the score distribution from the medical students

and the specialists (Jørgensen et al. 2018). The identified pass-fail score was used to explore the consequences of the identified standard: false positives (i.e. medical students who pass the test) and false negatives (i.e. specialists who fail the test) (Jørgensen et al. 2018).

Statistical analysis

Statistical analyses were performed using SPSS v. 23.0.0.0 (IBM Corp., Armonk, NY, USA) and G string III software (Papaworx, Hamilton, Ontario, Canada). Items were categorized into four levels (I, II, III, and IV) based on item difficulty (percentage of correctly answered items) and item discrimination using point biserial correlation statistics (Downing & Yudkowsky 2009). Internal consistency was explored by calculating Cronbach's α. Test scores between medical students and specialists were compared using independent samples t-test. P-values below 0.05 were interpreted as statistically significant. The intersect between the corresponding distribution curves were defined as the pass-fail score (contrasting groups' method) (Jørgensen et al. 2018). Based on Generalizability Theory and a Decision Study, we investigated how internal consistency measured as Generalizability coefficient would change when decreasing the number of MCQ items. This approach allows determining the minimal number of MCQ items needed for certification (>0.8).

Results

Content

A total of 100 MCQ items were developed based on the expert interviews. In the first Delphi-like iteration (100 items), 21 items were discarded, 52 items were rephrased based on comments from the experts, three new items were suggested, and 27 items were included in the final test without any changes. In the second Delphi-like iteration (55 items), four items were discarded, 22 new items were suggested, and 51 items were included in the final test. In the third Delphi-like iteration (22 items), 18 items were excluded and four items were included. This process yielded a total of 82 MCQ items **(Figure 1)**.

Response process

We recruited a total of 30 participants to take the developed test (20 medical students and 10 specialists in ophthalmology) **(Table 1)**. All had the test presented and supervised by the same instructor.

Internal structure

The items were categorized in four classification levels (I, II, III, and IV) **(Figure 2)**. This resulted in 42 items in level I (middle difficulty), 18 items in level II (easy), 9 items in level III (difficult) and 13 items in level IV (very easy or very difficult). Only items levels I and II were included in the final test (a total of 60 items), and the 22 level III and IV items were excluded. Final test with MCQ items (n=60) showed a very high level of internal consistency at Cronbach's $\alpha = 0.95$ (95% CI: 0.91 to 0.97).

Relationship to other variables

Relationship to other variables was investigated using the final test with 60 MCQ items. Each correct answer gave 1 point whereas incorrect answers gave 0 points. Hence, the overall test scores could range between 0 to 60. Medical students obtained a score of 30.0 ± 4.3 (mean ± standard deviation) (~50% correct on average). Specialists in ophthalmology obtained a score of 57.4 ± 1.6 (mean ± standard deviation) (~96% correct on average). These scores differed significantly ($P < 0.0001$, independent samples t-test). Because more specialists were females ($P = 0.05$, Exact test), we explored if differences between groups were gender-related and found no evidence of such ($P = 0.208$ for medical students, $P = 0.713$ for specialists in ophthalmology, when testing across genders using independent samples t-test).

Consequences

The score distributions of the medical students and the specialists intersect at 49.7 points (~50 points) (~83% correct), which represents the pass-fail point according to the contrasting groups' method **(Figure 3)**. At this pass-fail score, no medical students passed (false positive = 0%) and no specialists failed (false negative = 0%).

Sampling MCQs for a smaller test

We found that sampling a minimum of 15 of the 60 identified MCQ items is needed to obtain an internal consistency that is required for certification purposes **(Figure 4)**.

Discussion

In this study, we used a two-phased approach to develop and validate a theoretical test in direct ophthalmoscopy. We chose the MCQ format since it is argued that selected-response is the most suitable test format when testing knowledge and possesses the advantages of being low cost, the ability of testing large areas of knowledge in less time, and high reproducibility and reliability (Haladyna & Rodriguez 2013). Our final product from this study is a theoretical test with 60 MCQ items possessing strong validity evidence. We identified that sampling only 15 of these MCQ items suffice for certification purposes, which allows the 60 MCQ items to be used as a question bank.

The relevance of teaching medical students direct ophthalmoscopy has been discussed extensively (Yusuf et al. 2015, Purbrick & Chong 2015, Appleton & Nicholl 2016, Hill et al. 2016, Imonikhe et al. 2016). Critics argue that the procedure is too difficult, it is carried out too rarely, and when it is finally done, the quality of the procedure is poor, and the users confidence in their findings is low. Furthermore, there is little time for training in the undergraduates' curriculum (Purbrick & Chong 2015, Appleton & Nicholl 2016). Purbrick and Chong (2015) [51] suggest implementing fundus photography instead of teaching in the use of direct ophthalmoscopy. Others argue (Yusuf et al. 2015, Hill et al. 2016, Imonikhe et al. 2016) that performance of direct ophthalmoscopy has great diagnostic value and is a fundamental clinical skill that should be taught to all medical students. Immediate access to an ophthalmologist is not universal and clinical eye examination of a patient including direct ophthalmoscopy is essential in correct handling of the patient (Yusuf et al. 2015). The Principles and Guidelines of a Curriculum for Ophthalmic Education of Medical Students by the International Task Force on Ophthalmic Education of Medical Students (2006) of the International Council of Ophthalmology support this opinion and suggest teaching all medical students direct ophthalmoscopy.

Our expert panel for the first phase of the study had an overweight of experts in retinal diseases, which could be argued to influence the focus. Other ophthalmologists, or neurologists, emergency physicians, and family medicine practitioners might have provided other important insight. On the other hand, since examining the retina is the focus when using direct ophthalmoscopy, an overweight of experts in retinal diseases can also be considered very suitable. The selection of experts has been a point of debate when using the Delphi method. Critics argue that simply because individuals have knowledge of a specific area does not necessarily make them sufficient experts. Furthermore,

there is a potential of bias when selecting experts and the final panel may consist of the most willing experts and not necessarily the most competent (Graham et al. 2003). There is no gold standard describing how to select experts or how many experts are needed. A rule of thumb is that the reliability improves as the number of panelists increases (Graham et al. 2003). Streiner and Norman (2008) recommended 3-10 experts in the panel. We included five experts to the interviews and four of the experts attended in the Delphi-like process.

Our item analysis caused exclusion of 22 items that were either too difficult, too easy, or demonstrated poor ability to discriminate between proficiency levels. The final test consisted of 60 items that were either level I or II. These items demonstrated a very high level of internal consistency (Cronbach's $\alpha = 0.95$) well beyond the level required for certification purposes (> 0.8) (Downing 2004). Because the entire test has such a high level of internal consistency, it is possible to sample a smaller number of items (15 MCQ items) and still remain possess an internal consistency required for certification.

In this study, we used contrasting groups' method to identify a pass-fail score of 83% (Jørgensen et al. 2018). There are several methods to determine a pass-fail score (Downing & Yudkowsky 2009, Goldenberg et al. 2017); however, in a study aimed at setting pass scores for surgical tasks using Objective Structured Assessment of Technical Skill, De Montbrun et al. (2015) demonstrated that contrasting groups identify cut-off points at levels that are similar to those identified using other methods (i.e. borderline group and borderline regression) and provided evidence of consistency across the different methods. In contrast to contrasting groups' method, borderline-based methods require a group defined by being at the border of passing (i.e. 50% of the participants in the group should pass). Such a borderline group can be hard to identify, especially in a relatively unexplored field in direct ophthalmoscopy in terms of assessment and evaluation.

Important strengths and limitations should be noted. It is of upmost importance to follow contemporary best-practices when developing effective MCQs. In this study, we followed such practices that have been published by Case and Swanson (2001) and Haladyna and Rodriguez (2013). We diverged slightly from one recommended practice. Haladyna and Rodriguez (2013) suggest that when content experts are involved in item development (such as those in our phase I), they should first be thoroughly trained in item construction theories to better understand the notion of the feedback to

give. We did not find this endeavor to be feasible or realistic within our content experts' time schedule, so instead we trained the interviewer in item construction theories. Although this approach may yield results of acceptable quality, which we also show in this study; theoretically, if experts had been more aware of the end-product, they could have tailored their responses so that more or better questions could have been obtained. Evaluation of validity is an evaluation of whether a test is measuring what it is supposed to measure (Downing & Yudkowsky 2009). We followed the contemporary framework of validity described by Samuel Messick, which is a major strength of this study and an approach that is called for in the literature, which unfortunately is dominated by obsolete validity frameworks (Korndorffer et al. 2010, Borgersen et al. 2018). When using the contemporary framework by Samuel Messick, the validity evidence of a constructed test is evaluated from multiple sources, including content, response process, internal structure, relationship to other variables, and consequences. An evaluation from multiple sources is important and gives a comprehensive view of validity.

Considering the overall validity evidence, we suggest that our test can be used for measuring, evaluating, and certifying in theoretical aspects of direct ophthalmoscopy. This fills a gap in our current available toolset as educators, and we can now perform a range of interesting studies. For examples, interns who are expected to perform direct ophthalmoscopy can be screened for relevant knowledge, in a cheap and rapid manner. Or we can evaluate the yield of different educational initiatives using a more valid and reliable outcome measure. For example, one recent study found that participating in an art course significantly improved the observational skills in ophthalmology (Gurwin et al. 2018). Such interesting initiatives are currently very hard to evaluate and compare because of a lack of an outcome measure. This exact lack is addressed in our study. We have developed 60 well performing MCQ items on ophthalmoscopy knowledge **(Supplementary file)**. Subsets of these items (minimum 15) can be used to create tests for certification purposes or to compare different teaching methods and strategies as well as evaluation of new teaching methods. We advise inclusion of a test in standardized evidence-based training programs.

Acknowledgements

This study was supported by a grant from *Undervisningskvalitetspuljen* (a grant dedicated for quality improvements in medical education) from the University of Copenhagen. Authors MJ, MMS, CC, TB, JG, PBT, LK, TLS, and YS declare that they have no competing interests. Author FZ declare consulting fees unrelated to this work from Alimera, Allergan, Bayer HealthCare, Novartis, MSD, Roche and speaker fees unrelated to this work from Alcon, Alimera, Allergan, Bayer HealthCare, and Novartis.

References

Appleton JP & Nicholl DJ (2016): Comment on: 'Direct ophthalmoscopy should be taught to undergraduate medical students'. Eye (Lond) **30**: 327.

Bacon AS, Dart JK, Ficker LA, Matheson MM & Wright P (1993): Acanthamoeba keratitis. The value of early diagnosis. Ophthalmology **100**: 1238-43.

Bek T, Hjortdal J & La Cour M (2016): [Eye diseases]. 2nd edition. FADLs Forlag.

Bloch R & Norman G (2012): Generalizability theory for the perplexed: a practical introduction and guide: AMEE Guide No. 68. Med Teach **34**: 960-992.

Borgersen NJ, Naur TMH, Sørensen SMD, Bjerrum F, Konge L, Subhi Y & Thomsen ASS (2018): Gathering Validity Evidence for Surgical Simulation: A Systematic Review. Ann Surg [ePub ahead of print 4/Jan/2018] doi: 10.1097/SLA.0000000000002652.

Bruce BB, Lamirel C, Wright DW, Ward A, Heilpern KL, Biousse V & Newman NJ (2011): Nonmydriatic ocular fundus photography in the emergency department. N Engl J Med **364**: 387-389.

Case SM & Swanson DB (2001): Constructing written test questions for the basic and clinical sciences. 3rd edition. National Board of Medical Examiners.

Chawla B, Hasan F, Azad R, Seth R, Upadhyay AD, Pathy S & Pandey PM (2016): Clinical presentation and survival of retinoblastoma in Indian children. Br J Ophthalmol **100**: 172-178.

De Montbrun S, Satterthwaite L & Grantcharov TP (2015): Setting pass scores for assessment of technical performance by surgical trainees. Br J Surg **103**: 300-306

Downing SM & Yudkowsky R (eds.) (2009): Assessment in health professions education. 1st edition. Routledge.

Eijk ES, Busschbach JJ, Timman R, Monteban HC, Vissers JM & van Meurs JC (2016): What made you wait so long? Delays in presentation of retinal detachment: knowledge is related to an attached macula. Acta Ophthalmol **94**: 434-440.

Fahmy P, Hamann S, Larsen M & Sjølie AK (2013): [Practical ophthalmology]. 3rd edition. Gads Forlag.

Georgouli T, Pountos I, Chang BY & Giannoudis PV (2011): Prevalence of ocular and orbital injuries in polytrauma patients. Eur J Trauma Emerg Surg **37**: 135-140.

Goldenberg MG, Garbens A, Szasz P, Hauer T & Grantcharov TP (2017): Systematic review to establish absolute standards for technical performance in surgery. Br J Surg **104**: 13-21.

Graham B, Regehr G & Wright JG (2003): Delphi as a method to establish consensus for diagnostic criteria. J Clin Epidemiol **56**: 1150-1156.

Gupta RR & Lam WC (2006): Medical students' self-confidence in performing direct ophthalmoscopy in clinical training. Can J Ophthalmol **41**: 169-174.

Gurwin J, Revere KE, Niepold S, Bassett B, Mitchell R, Davidson S, DeLisser H & Binenbaum G (2018): A randomized controlled study of art observation training to improve medical student ophthalmology skills. Ophthalmology **125**: 8-14.

Haladyna TM & Rodriguez MC (eds.) (2013): Developing and validating test items. 1st edition. Routledge.

Harper RA (2016): Basic Ophthalmology, essentials for medical students. 10th edition. American Academy of Ophthalmology.

Hill SC, Jawais I & Amoaku W (2016): Response to: 'Direct ophthalmoscopy should be taught to undergraduate medical students'. Eye (Lond) **30**: 327-328.

Imonikhe RJ, Finer N, Gallagher K, Plant G, Bremner FD & Acheson JF (2016): Direct ophthalmoscopy should be taught to undergraduate medical students. Eye (Lond) **20**: 497.

International Task Force on Ophthalmic Education of Medical Students (2016): Principles and guidelines of a curriculum for ophthalmic education of medical students. Klin Monbl Augenheilkd **223 Suppl 5**: S1-19.

Jensen JT, Savran MM, Møller AM, Vilmann P, Hornslet P & Konge L (2016): Development and validation of a theoretical test in non-anaesthesiologist-administered propofol sedation for gastrointestinal endoscopy. Scand J Gastroenterol **51**: 872-879.

Joint American Educational Research Association (2014): The Standards for Educational and Psychological Testing. American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME).

Jørgensen M, Konge L & Subhi Y (2018): Contrasting groups' standard setting for consequences analysis in validity studies: reporting considerations. Adv Simul (Lond) **3**: 5.

Kohls-Gatzoulis JA, Regehr G & Hutchison C (2004): Teaching cognitive skills improves learning in surgical skills courses: a blinded, prospective, randomized study. Can J Surg **47**: 277-283.

Korndorffer JR, Kasten SJ & Downing SM (2010): A call for the utilization of consensus standards in the surgical education literature. Am J Surg **199**: 99-104.

Kromann CB, Jensen ML & Ringsted C (2009): The effect of testing on skills learning. Med Educ **43**: 21-27.

Leske MC, Heijl A, Hussein M, Bengtsson B, Hyman L & Komaroff E (2003): Factors for glaucoma progression and the effect of treatment: the early manifest glaucoma trial. Arch Ophthalmol **121**: 48-56.

Mackay DD, Garza PS, Bruce BB, Newman NJ & Biousse V (2015): The demise of direct ophthalmoscopy: A modern clinical challenge. Neurol Clin Pract **5**: 150-157.

Morad Y, Barkana Y, Avni I & Kozer E (2004): Fundus anomalies: what the pediatrician's eye can't see. Int J Qual Health Care **16**: 363-365.

Nicholl DJ, Yap CP, Cahill V, Appleton J, Willetts E & Sturman S (2012): The TOS study: can we use our patients to help improve clinical assessment? J R Coll Physicians Edinb **42**: 306-310.

Patil SG, Kotwal IA, Joshi U, Allurkar S, Thakur N & Aftab A (2016): Ophthalmological Evaluation by a Maxillofacial Surgeon and an Ophthalmologist in Assessing the Damage to the Orbital Contents in Midfacial Fractures: A Prospective Study. J Maxillofac Oral Surg **15**: 328-335.

Purbrick RM & Chong NV (2015): Direct ophthalmoscopy should be taught to undergraduate medical students – no. Eye (Lond) **29**: 990-991.

Rasmussen A, Brandi S, Fuchs J, Hansen LH, Lund-Andersen H, Sander B & Larsen M (2015): Visual outcomes in relation to time to treatment in neovascular age-related macular degeneration. Acta Ophthalmol **93**: 616-620.

Ringsted CV, Pallisgaard J & Falck G (2010): [Physicians' clinical skills after finishing internship]. Ugeskr Laeger **164**: 3211-3215.

Savran MM, Clementsen PF, Annema JT, Minddal V, Larsen KR, Park YS & Konge L (2014): Development and validation of a theoretical test in endosonography for pulmonary diseases. Respiration **88**: 67-73.

Savran MM, Hansen HJ, Petersen RH, Walker W, Schmid T, Bojsen SR & Konge L (2015): Development and validation of a theoretical test of proficiency for video-assisted thorascopic surgery (VATS) lobectomy. Surg Endosc **29**: 2598-2604.

Schuwirth LW & van der Vleuten CP (2003): ABC of learning and teaching in medicine: Written assessment. BMJ **326**: 643-645.

Streiner DL & Norman GR (eds) (2008): Health Measurement Scales: a Practical Guide to Their Development and Use. 4th edition. Oxford University Press.

Thomsen AS, Subhi Y, Kiilgaard JF, la Cour M & Konge L (2015): Update on simulation-based surgical training and assessment in ophthalmology: a systematic review. Ophthalmology **122**: 1111-1130.

408 Ting DS, Sim SS, Yau CW, Rosman M, Aw AT & Yeo IY (2016): Ophthalmology simulation for
409 undergraduate and postgraduate clinical education. Int J Ophthalmol **9**: 920-924.

410 Yusuf IH, Salmon JF & Patel CK (2015): Direct ophthalmoscopy should be taught to undergraduate
411 medical students – yes. Eye (Lond) **29**: 987-989.

412

Tables

Table 1. Characteristics of the participants who took the developed test to collect validity evidence.

	Medical students (n = 20)	Specialists in ophthalmology (n = 10)
Age, years, mean (SD)	23 (3)	46 (8)
Females, n (%)	7 (35)	8 (80)
Previous experience*, median (IQR)	0 (0 to 0)	1,000 (200 to 2,000)

Abbreviations: SD = standard deviation; IQR = interquartile range.
*: Previous experience was defined as number of ophthalmoscopies performed previously.

Figure legends

Figure 1. Flowchart of the Delphi-like process in which the content of the test was determined through content expert consensus on phrasing and relevance. A total of 82 MCQs were included for further evaluation.

Figure 2. Item analysis. Horizontal axis: Item discrimination; the higher item discrimination, the higher discrimination ability of the item. Vertical axis: Item difficulty; refers to the proportion of examinees who answered an item correctly (an item difficulty of 1,00 means that 100% of the examinees answered the item correctly). Item class: All items were categorized in four levels based on item discrimination and item difficulty. Level I (middle difficulty and high discriminatory ability), Level II (easier difficulty with high discriminatory ability), Level III (too difficult), Level IV (too easy or too low discriminatory ability). Level I and II items were included in the final test.

Figure 3. The contrasting groups' method is used for consequence analysis to establish a pass-fail score. The score curves from the medical students (green) and the specialists (blue) are used to identify the intersection point (black vertical line) and thereby the pass-fail score (49.7 ~ 50 points). The intersect is shown at a larger zoom in the red box.

Figure 4. Generalizability coefficient for increasing number of MCQ items obtained through a Decision study allows exploring whether sampling a smaller number of MCQs is possible while still possessing an internal consistency required for certification. Stripes show that sampling only 15 MCQ items suffice.