Reliable test of clinicians' mastery in skin cancer diagnostics

Ternov, Niels Kvorning; Vestergaard, T.; Hölmich, L. Rosenkrantz; Karmisholt, K.; Wagenblast, A. L.; Klyver, H.; Hald, M.; Schøllhammer, L.; Konge, L.; Chakera, A. H.

Published in:
Archives of Dermatological Research

DOI:
10.1007/s00403-020-02097-8

Publication date:
2021

Document version:
Accepted manuscript

Citation for pulished version (APA):
Ternov, N. K., Vestergaard, T., Hölmich, L. R., Karmisholt, K., Wagenblast, A. L., Klyver, H., Hald, M., Schøllhammer, L., Konge, L., & Chakera, A. H. (2021). Reliable test of clinicians' mastery in skin cancer diagnostics. *Archives of Dermatological Research*, *313*(4), 235-243. https://doi.org/10.1007/s00403-020-02097-8

Reliable Test of Clinicians' Mastery in Skin Cancer Diagnostics

Key words: Continuous Medical Education, Skin Cancer, Melanoma, Nevi
Short title: **Skin cancer MCQ**
Word Count: 2999
Table Count: 1
Figures: 5
Tables: 1

Authors:

N. K. Ternov[1,8], MD	niels.kvorning.ternov@regionh.dk
T. Vestergaard[2,9], MD	tine.vestergaard@rsyd.dk
L. Rosenkrantz Hölmich[1,8], MD, MSCE	lisbet.rosenkrantz.hoelmich@regionh.dk
K. Karmisholt[3], MD, PhD	katrine.elisabeth.karmisholt@regionh.dk
A.L.Wagenblast[4], MD	anne.lene.wagenblast@regionh.dk
H. Klyver[4], MD	helle.klyver@regionh.dk
M. Hald[5,8], MD, PhD	marianne.hald@regionh.dk
L. Schøllhammer[6], MD	liv.schollhammer@rsyd.dk
L. Konge[7,8], MD, PhD	lars.konge@regionh.dk
A. H. Chakera[1,8], MD, PhD	annette.hougaard.chakera@regionh.dk

[1]Department of Plastic Surgery, Herlev and Gentofte University Hospital, Copenhagen, Denmark
[2]Department of Dermatology and Allergy Centre, Odense University Hospital, Odense, Denmark
[3]Department of Dermatology, Bispebjerg University Hospital, Copenhagen, Denmark
[4]Department of Plastic Surgery, breast surgery and Burns Treatment, Copenhagen University Hospital, Rigshospitalet, Copenhagen, Denmark
[5]Department of Dermatology, Herlev and Gentofte University Hospital, Copenhagen, Denmark
[6]Department of Plastic Surgery, Odense University Hospital, Odense, Denmark
[7]Copenhagen Academy for Medical Education and Simulation, Copenhagen, Denmark
[8]Faculty of Health and Medical Sciences, Copenhagen University, Copenhagen, Denmark
[9]Faculty of Health Sciences, University of Southern Denmark, Copenhagen, Denmark

Corresponding Author:
Niels Kvorning Ternov
Niels.kvorning.ternov@regionh.dk
ORCID: 0000-0002-8034-6727
Department of Plastic Surgery
Herlev and Gentofte University Hospital
Herlev Ringvej 75
2730 Herlev
Denmark

Abstract

Background:
Differentiating between benign and malignant skin lesions can be very difficult and should only be done by sufficiently trained and skilled clinicians. To our knowledge there are no validated tests for reliable assessments of clinicians' ability to perform skin cancer diagnostics.

Objective:
To develop and gather validity evidence for a test in skin cancer diagnostics.

Methods:
A multiple-choice questionnaire (MCQ) was developed based on informal interviews with seven content experts from five skin cancer centers in Denmark. Validity evidence for the test was gathered from May until July 2019 using Messick's validity framework (content, response process, internal structure, relationship to other variables and consequences). Item *content* was revised through a Delphi-like review process and then piloted on 36 medical students and 136 doctors using a standardized *response process*. Results enabled an analysis of the *internal structure* and *relationship to other variables* of the test. Finally, the contrasting groups method was used to investigate the test's *consequences (pass-fail standard)*.

Results:
The initial 90-item MCQ was reduced to 40 items during the Delphi-like review process. Item analysis revealed that 25 of the 40 selected items were level I-III quality items with a high internal consistency (Cronbach's alpha = 0.83) and highly significant ($P \leq 0.0001$) differences in test scores between participants with different occupations or levels of experience. A pass-fail standard of 12 (48%) correct answers was established using the contrasting groups' method.

Conclusion:
The skin cancer diagnostics MCQ developed in this study can be used for reliable assessments of clinicians' competencies.

Declarations

Funding: Salary (5620 €) for the primary investigator was funded by Tvaerspuljen (Capital Region of Denmark) a government grant. The funder had no influence on the study design, data collection, data analysis, manuscript preparation or publication decisions.

Conflicts of interest: Dr. Ternov, Dr. Hölmich and Dr. Chakera reports grants from Region Hovedstaden, during the conduct of the study. Dr. Vestergaard, Dr. Karmisholt, Dr. Wagenblast, Dr. Hald, Dr. Klyver, Dr. Schøllhammer and Dr. Konge has nothing to disclose.

Availability of data: All data is available upon request.

Background

Skin cancer is the most prevalent cancer among fair skinned people worldwide and the incidence of melanoma has increased by 2-7% annually for more than 30 years [1]. Future prognoses predict an additional 100% increase by the year 2030 [2]. Although there have been significant advances in the treatment of metastatic melanoma within the recent decade [3], early diagnosis remains the most important prognostic factor for melanoma survival [4].

National guidelines from the UK, Germany, Australia, New Zealand, Denmark recommend that skin lesions suspicious of melanoma should be diagnosed based on a brief medical history and unaided visual as well as dermoscopic evaluation [5–9]. A recent meta-analysis found that this method for melanoma detection yields a sensitivity and specificity of 92.4% and 79.7%, respectively [10]. Dermoscopy increases the odds of diagnosing skin lesions correctly by 49% compared to an unaided visual examination [11].

This increased accuracy is, however, only possible when clinicians have obtained a sufficient level of competency in dermoscopy, which requires extensive training [12, 13]. Studies from the US, UK, Australia, Austria, and Canada found that 30-86% of residents in dermatology were unsatisfied with their training in dermoscopy and up to 40% felt they lacked confidence in dermoscopy of skin lesions [14–18].

Clinicians have expressed concerns regarding these findings and recommend specific training courses and competency testing in dermoscopy within the curriculum for dermatology training programs [19]. Introducing tests in the curriculum is the only way to ensure uniform competence

and might in itself increase the retention of knowledge through the so-called testing effect [20, 21].

Careful considerations are required when deciding upon the type, format, content, validity, reliability, and cost-effectiveness of a test to be incorporated in official training programs [22–24]. Multiple choice questionnaires (MCQs) are known to be reproducible, reliable, and offer time- and cost-effective assessment of trainees' knowledge within an extensive syllabus [24]. However, it is essential to gather specific validity evidence for every newly developed test to prove that the test measures what it is supposed to [23].

MCQ tests for melanoma diagnostics have previously been described, however, to the best of our knowledge, none of them have been developed and validated using best practices for MCQ development and a modern validity framework [25–28]. Thus, the aim of this study was to develop and gather validity evidence for a MCQ test for skin cancer diagnostics, following best practice guidelines for MCQ development and Messick´s validity framework [23].

Methods

This study included two consecutive phases; developing the test and gathering validity evidence. In the first phase, we developed test items consisting of a broad range of photos and related clinical information of benign and malignant skin lesions. Choice of skin lesion photos was based on consensus from informal interviews with seven clinical content experts from five skin cancer centers in Denmark; three consultant dermatologists (TV, KK, and MH) and four consultant plastic surgeons (AC, LH, AW, and HK). All experts had more than 6 years of experience diagnosing skin lesions and were clinically active during the time of participation.

During the second phase, we collected validity evidence for the MCQ. We initially reviewed the photos and clinical content through a Delphi-like process, utilizing the knowledge of the seven experts. The Delphi method is an established method for systematic anonymous consensus agreement among a panel of experts [29]. We then tested the content on recruited test subjects with varying levels of proficiency ranging from 0 to >10 years of experience in skin lesion diagnostics. The included subjects were medical students and doctors employed at clinics for general medicine and departments of dermatology or plastic surgery. All medical students had passed their exams in dermatology and thus had a conceptual understanding of skin lesion diagnostics although they had no clinical experience.

The study was conducted in accordance with the Declaration of Helsinki. All participants were informed about the study prior to participation and gave informed consent. The study held no consequences for the participants and all collected data was fully anonymous. Test participation was voluntary and according to Danish legislation the study was exempt from ethical approval per the Ethics Committee of the Capital Region of Denmark.

Phase I: Test development

We initially reviewed the literature, looking specifically for factors of importance for skin lesion diagnostics and found the following; patient age, location of the skin lesion, visual examination with and without dermoscopy, palpation and a focused medical history [5–9, 30, 31] .

After reviewing the literature, the primary investigator (NT) held informal interviews concerning test content, with all members of the expert panel. Answers were noted and later summarized in a report sent out by email to all seven experts. Final consensus was reached through email correspondence.

Item content was constructed according to the MCQ guidelines by Case & Swanson [22]. The selected item format was one-best-answer questions with a stem consisting of skin lesion cases with a lead-in-question and a series of seven answer options with one best answer, 5 misleading options and an "I have no clue" option. The lead-in-question and answering options were identical and placed in the same order for each item to avoid confusion and potential bias from testwiseness [22].

The skin lesion cases used during development of the MCQ were extracted from the photo library and electronic patient journals at the Department of Dermatology and Allergy Centre, Odense University Hospital. Permission to access and handle the patients' data was granted by the Danish Patient Safety Authority (Jr. Nr. 3-3013-2553/1) and the Data Protection Agency of Southern Denmark (Jr. Nr. 18/53664). All skin lesion photos and descriptions were anonymized before inclusion in the test. The diagnosis of each case was either verified through histopathology or clinical consensus among a minimum of eight board approved dermatologists and plastic surgeons.

Phase II: Test validity

Validity evidence for the MCQ items was gathered using the framework by Samuel Messick, stressing the importance of test content, response process, internal structure, relationship to other variables and consequence [23].

Content – Relevance of content for the intended use

The initial draft of the MCQ developed during "Phase I" was assessed by all seven experts in a Delphi-like review process [32] consisting of two consecutive rounds (Fig. 1). During the first

review round, experts were asked to rate the items in terms of quality and clinical relevance on two separate five-step Likert scales [33] whilst simultaneously editing/commenting upon related clinical text content. Items with a mean score in relevance or quality below 4 were removed from the test. Selection limits were not revealed to experts to avoid bias. During the second review round, redundant or ambiguous items were removed. Items were categorized as redundant or ambiguous if there were multiple similar lesions in the test or experts were unable to reach a consensus concerning the diagnosis. The final test items were piloted on two medical students and two clinicians, ensuring that the test was straightforward and easy to understand.

Response process – Eliminating or controlling response bias

Medical students were recruited after a lecture unrelated to skin lesion diagnostics. Doctors employed at clinics for general medicine were recruited during a meeting for residents partaking in the last rotation of their residency. Remainder clinicians were recruited during clinical conferences at the departments of dermatology and plastic surgery at the five included university hospitals. All included subjects were supervised by the first author before and during participation in the MCQ test. The seven experts who participated in the test development and content validation were excluded from participation.

Participants answered the test through Google Forms (Google, Mountain View, California, USA) using their smartphones.

Internal structure – Quality of items and test reliability

Power calculations could not be performed beforehand on this newly developed test. However, our goal was to develop a test that enabled valid assessments of the various levels of proficiency and therefore >10 participants were included per proficiency group, to achieve normal distribution

[34]. Item difficulty index (DI) and item discrimination index (DsI) were calculated based on participants' test results. The DI is the proportion of participants answering each item correctly and the DsI corresponds to an item's ability to differentiate between test subjects with a high and low overall performance, respectively.

Items were categorized into four established levels[35]: Level **I** (intermediate difficult, DI = 0.45-0.75, DsI > 0.20), Level **II** (easy, DI = 0.76-0.91, DsI > 0.15), Level **III** (difficult, DI = 0.25-0.44, DsI > 0.10) and Level **IV** (very easy/difficult, DI < 0.25 or > 0.91, respectively, any DsI). Level I items are good, level II and III items are acceptable and level IV and remaining items are considered inappropriate for use. For this reason, we decided to only include level I-III items. The overall reliability of level I-III items was tested to estimate the internal consistency of the MCQ, i.e. that all the included items related to assessment of the same overall competence (skin cancer diagnostics).

Relationship to other variables – Correlation between test scores and external variables expected to influence the test results (e.g. experience or occupation/specialty)

We evaluated the relationship between test results and clinicians' cumulative experience diagnosing skin lesions as well as occupation/specialty (medical student, general medicine, plastic surgery, and dermatology). Both experience and occupation/specialty were expected to influence the final test scores, rendering better score results for those with an extensive experience and/or an employment at specialized centers (e.g. departments of dermatology).

Consequence – Consequence from obtained test scores

A pass-fail standard was established using the contrasting groups' method [36] defining novices as those with no experience and experts as those with >6 years of experience diagnosing skin

lesions. A good discriminatory ability is essential for a clinically useful test. Thus, the pass/fail-standard should result in few false positives and false negatives, i.e. novices passing and experts failing the test.

Statistical analysis

Statistical analyses were performed using SPSS v. 25.0.0.0 (IBM Corp., Armonk, NY, USA) and G string III software (Papaworx, Hamilton, Ontario, Canada). The DI and DsI were calculated using point biserial correlation statistics [35]. The internal consistency reliability was calculated using Cronbach's α. Generalizability theory including a Decision Study was applied to identify the minimum number of items needed for acceptable (Generalizability coefficient > 0.7) or good (Generalizability coefficient > 0.8) reliability. Test results from the various experience levels and occupation/specialty groups were compared using an ANOVA test with Bonferroni's correction for multiple comparisons. Multivariable regression analysis was used to evaluate the relationship between occupation and test results. P-values of <0.05 were considered statistically significant. A pass-fail limit was calculated as the point of intersection between the normal distribution graphs for test results by novices and experts.

Results

Phase I: Test development

Expert interviews yielded a consensus regarding item construction; a stem consisting of a focused medical history and photos (clinical and dermoscopic), a lead-in-question ("What is the diagnosis?") and seven answering options ("I have no clue", "Nevus", "Melanoma or melanoma in situ", "Seborrheic keratosis or solar lentigo", "Carcinoma (basal or squamous cell)", "Dermatofibroma", "Vascular tumor" and "Other (benign) lesion").

Phase II: Test validity

Content

The initial test draft based on informal interviews included 90 items of which 30 and 20 were excluded during the first and second review rounds, respectively (Fig. 1). Thus, the final test draft consisted of 40 items.

[Insertion of Fig. 1]

Response process

We recruited a total of 172 participants at various experience levels; 36 medical students and 136 doctors (general medicine (n=54), plastic surgery (n=51), and dermatology (n=31)). All participants answered the MCQ with 40 items under supervision by the primary investigator during the period May-July 2019.

Internal structure

Item analysis revealed a mean DI and DsI of 0.5 and 0.27, respectively. There were 10 level I, 3 level II, 10 level III, 6 level IV and 11 remainder items (Fig. 2). The quality of six level IV and 9 of the remainder items were too bad to include in the final test. However, two of the remainder items had good DIs (0.45-0.75) and acceptable DsIs (0.15-0.20) and were thus classified as level IB items (orange with red stripes, Fig. 2). This resulted in a final test consisting of 25 level I-III items with a high internal consistency reliability (Cronbach's alpha = 0.83). Mean DI and DsI of the 25 level I-III items were 0.51 and 0.36, respectively, and constituted: 5 melanomas/melanoma in situ, 2 vascular tumors, 8 nevi, 2 basal cell carcinomas, 5 seborrheic keratoses, 1 dermatofibroma, 1 subungual melanoma, and 1 subungual hemorrhage.

[Insertion of Fig. 2]

Relationship to other variables

The maximum test score was 25 points (1 point per item). Table 1 shows the mean, minimum and maximum test scores, standard deviations and confidence intervals for the various occupation/specialty groups and levels of experience. Fig. 3 shows the median, minimum, maximum, range and 95% confidence interval of test scores at various levels of experience. A one-way ANOVA test showed significant intergroup variance between participants' proficiency both when divided by experience ($F = 21.01$, $p < 0.001$) and occupation/specialty ($F = 102.60$, $p < 0.001$).

Doctors with ≥6 years of experience performed significantly better than those with <6 years of experience ($p \leq 0.001$) (Table 1). Our results indicate a very slow learning curve among doctors as there were no significant differences in the performance of doctors with levels of experience ranging from 0-3 months to 5 years (Fig. 3). Furthermore, doctors employed at departments of dermatology performed significantly better than remainder occupation/specialty groups.

Through a multivariable regression analysis on test results from clinicians employed at departments of plastic surgery and dermatology we found that experience was a significant predictor for outcome ($r = 0.35$, adjusted $R^2 = 0.14$, $p = 0.001$). However, when occupation/specialty was added, the model for predicting test results became significantly better ($r = 0.49$, adjusted $R^2 = 0.23$, $p = 0.001$), as clinicians from departments of dermatology performed significantly better than those employed at departments of plastic surgery, irrespective of the included clinicians' level of experience.

[Insertion of Fig. 3 and Table 1]

Consequence

The normal distribution for test scores among novices and experts intersected at 12 out of 25 correct answers (48%). At this pass-fail limit one novice (3.1%) passed and two experts (4.9%) failed.

[Insertion of Fig. 4]

Decision study

The decision study showed that 12 randomly sampled items resulted in an acceptable reliability (Generalizability coefficient > 0.7) for low-stakes tests (e.g. to provide feedback to trainees after a course). Twenty randomly selected items were necessary to provide a sufficient reliability (Generalizability coefficient > 0.8) for a high stakes test, e.g. for certification (Fig. 4).

[Insertion of Fig. 5]

Discussion

Summary of findings

In this study, we developed and found strong validity evidence for a 25-item MCQ (Appendix 1) for objective evaluation of a clinician's mastery in skin cancer diagnostics. Item content was thoroughly reviewed by a panel of seven content experts utilizing a Delphi-like review process and we found a high reliability (Cronbach's alpha = 0.83) of the final test, meaning that the included items measured the same overlying construct.

As expected we found a highly significant correlation between participants' test results and their level of experience (P<0.0001), which provides additional evidence that the MCQ in fact does test the targeted area (skin cancer diagnostics). Furthermore, a Decision Study (Fig. 4) showed that items can be randomly sampled among the 25 items to create either smaller 12-item tests with an acceptable reliability (Generalizability coefficient >0.7) for feedback during training or more

comprehensive 20-item tests with a good reliability (Generalizability coefficient >0.8) for more formal assessments. Finally, through the contrasting groups' method we established a pass-fail limit of 12/25 correct answers.

Limitations

The initial MCQ items were developed after conducting semi-structured and informal interviews with seven experts, inspired by former studies [36–38]. The chosen number of experts (n=7) was based on recommendations from the literature [39, 40]. Unfortunately, there is no consensus regarding the definition of expert knowledge within the literature [39, 41, 42]. Thus, without supporting evidence, we defined experts as clinically active board certified medical specialists in dermatology or plastic surgery with >6 years of experience diagnosing skin lesions.

Experts are necessary for content development and the Delphi approach prevents subject bias from peer group pressure [29]. However, the general foundation of Delphi is subjectivity [43] and thus we cannot guarantee that there is no bias associated with our choice of experts and their inherent opinions.

It is recommended that content experts are thoroughly trained in item construction before participation. However for practical reasons, we only trained the primary investigator, as has been done in multiple similar studies [37, 38, 44].

Some may argue that the specific diagnosis is less important than the potential malignancy of a lesion, i.e. whether the lesion is "malignant" or "benign". However, it is strongly recommended that one-true-answer questions include a minimum of three options, otherwise the average score would be 50% when test answers are chosen randomly [22].

We included two level IB items within the final MCQ, as their DIs were within the optimal range (0.45-0.75) whilst their discriminative abilities were acceptable but slightly lower than generally recommended (DsI=0.15-20) [23]. Inclusion of the two level IB items increased the reliability and coverage of the MCQ test whilst only marginally reducing the mean DsI from 0.377 to 0.360.

It should be noted that 11 out of our 25 items were difficult (DI<0.44), reflected in the low pass-fail limit 12/25 (48%) found during consequence analysis.

Implications

There is an unmet need for efficient educational interventions in skin cancer diagnostics and reliable tests for objective assessment of clinicians' level of mastery. This is further supported by the very long learning curve we found in this study, which seems to extend beyond five years of experience.

Some may argue that recent advances in computerized diagnostics through image recognition by convolutional neural networks (CNN) [45, 46] will render clinical examinations and the training of clinicians therein obsolete. Nevertheless, our beliefs coincide with the international consensus that CNNs will mainly aid clinicians through decisional support, whilst the primary and final diagnostics will continue to be performed by clinicians [47, 48]. Whilst future clinical implementation of validated CNN algorithms may save time and provide more efficient and safe skin lesion diagnostics, they will also increase the need for highly trained clinicians providing the final diagnosis and treatment strategy based on both algorithmic and clinical input, especially in difficult and ambiguous cases. It is therefore essential that clinicians receive efficient training in skin lesion diagnostics and that their competencies are tested using reliable tests.

Tables

				95% CI				Comparison of means	
	Total(n)	Mean	SD	Lower limit	Upper limit	Min.	Max.	Mean Δ	P-value
Experience:									
None	32	7.6	2.2	6.8	8.4	3	12	-10.9	<0.001
0-3 months	13	9.8	4.7	6.9	12.6	2	17	-8.7	<0.001
4-11 months	16	11.1	4.0	9.0	13.2	6	19	-7.4	<0.001
1-2 years	37	13.0	4.5	11.5	14.5	4	22	-5.5	0.001
3-5 years	33	12.5	5.7	10.4	14.5	3	24	-6.0	<0.001
6-10 years	13	18.5	2.6	16.9	20.0	14	22	0.0	-
>10 years	28	17.7	3.1	16.5	18.9	11	23	-0.8	1.00
Occupation/ Specialty:									
Medical student	35	7.5	2.5	6.6	8.3	2	12	-10.9	<0.001
General medicine	54	9.4	2.9	8.6	10.2	3	16	-9.0	<0.001
Plastic surgery	51	16.0	3.5	15.1	17.0	9	22	-2.4	0.007
Dermatology	32	18.4	3.8	17.1	19.8	9	24	0.0	-

Table 1 - Mean/min/max scores, standard deviations, 95% confidence intervals and difference in means (including significance) for participant groups divided both by level of experience and occupation/employment.

References

1. Whiteman DC, Green AC, Olsen CM (2016) The Growing Burden of Invasive Melanoma: Projections of Incidence Rates and Numbers of New Cases in Six Susceptible Populations through 2031. J Invest Dermatol 136:1161–1171 . https://doi.org/10.1016/j.jid.2016.01.035
2. Guy GP, Machlin SR, Ekwueme DU, Yabroff KR (2015) Prevalence and Costs of Skin Cancer Treatment in the U.S., 2002−2006 and 2007−2011. Am J Prev Med 48:183–187 . https://doi.org/10.1016/j.amepre.2014.08.036
3. Domingues B, Lopes J, Soares P, Populo H (2018) Melanoma treatment in review. ImmunoTargets Ther Volume 7:35–49 . https://doi.org/10.2147/ITT.S134842
4. Survival Rates for Melanoma Skin Cancer. https://www.cancer.org/cancer/melanoma-skin-cancer/detection-diagnosis-staging/survival-rates-for-melanoma-skin-cancer-by-stage.html. Accessed 16 Apr 2019
5. Vissing S, Hald M, Mosegaard M, Fabricius S, Vestergaard T Kliniske guidelines vedrørende undersøgelse, diagnostik og behandling af kutane melanocytære nævi og kutant malignt melanom (MM). https://dds.nu/wp-content/uploads/2019/04/MM-guideline-final-til-DDS-2019-ok.pdf. Accessed 25 Nov 2019
6. NICE (National Institute for Health and Care Excellence) (2015) Melanoma: assessment and management. https://www.nice.org.uk/guidance/ng14/chapter/1-Recommendations#assessing-melanoma-2. Accessed 25 Aug 2019
7. Scottish Intercollegiate Guidelines Network (SIGN) (2017) Cutaneous Melanoma. A national clinical guideline. https://www.sign.ac.uk/sign-146-melanoma.html. Accessed 25 Aug 2019
8. Leitlinienprogramm Onkologie (2019) S3-Leitlinie zur Diagnostik, Therapie und Nachsorge des Melanoms, Version 3.2. https://www.awmf.org/uploads/tx_szleitlinien/032-024OLl_S3_Melanom-Diagnostik-Therapie-Nachsorge_2019-11_1.pdf. Accessed 25 Nov 2019
9. Clinical practice guidelines for the management of melanoma in Australia and New Zealand. In: Minist. Health NZ. https://www.health.govt.nz/publication/clinical-practice-guidelines-management-melanoma-australia-and-new-zealand. Accessed 25 Aug 2019
10. Dinnes J, Deeks JJ, Grainge MJ, Chuchu N, Ferrante di Ruffano L, Matin RN, Thomson DR, Wong KY, Aldridge RB, Abbott R, Fawzy M, Bayliss SE, Takwoingi Y, Davenport C, Godfrey K, Walter FM, Williams HC, Cochrane Skin Cancer Diagnostic Test Accuracy Group (2018) Visual inspection for diagnosing cutaneous melanoma in adults. Cochrane Database Syst Rev 12:CD013194 . https://doi.org/10.1002/14651858.CD013194
11. Kittler H, Pehamberger H, Wolff K, Binder M (2002) Diagnostic accuracy of dermoscopy. Lancet Oncol 3:159–165 . https://doi.org/10.1016/S1470-2045(02)00679-4
12. Patel P, Khanna S, McLellan B, Krishnamurthy K (2017) The need for improved dermoscopy training in residency: a survey of US dermatology residents and program directors. Dermatol Pract Concept 7:17–22 . https://doi.org/10.5826/dpc.0702a03
13. Terushkin V, Warycha M, Levy M, Kopf AW, Cohen DE, Polsky D (2011) Analysis of the Benign to Malignant Ratio of Lesions Biopsied by a General Dermatologist Before and After the Adoption of Dermoscopy. Arch Dermatol 146: . https://doi.org/10.1001/archdermatol.2010.12
14. Wu TP, Newlove T, Smith L, Vuong CH, Stein JA, Polsky D (2013) The importance of dedicated dermoscopy training during residency: A survey of US dermatology chief residents. J Am Acad Dermatol 68:1000–1005 . https://doi.org/10.1016/j.jaad.2012.11.032
15. Terushkin V, Oliveria SA, Marghoob AA, Halpern AC (2010) Use of and beliefs about total body photography and dermatoscopy among US dermatology training programs: An update. J Am Acad Dermatol 62:794–803 . https://doi.org/10.1016/j.jaad.2009.09.008
16. Butler TD, Matin RN, Affleck AG, Fleming CJ, Bowling JC (2015) Trends in dermoscopy use in the UK: results from surveys in 2003 and 2012. Dermatol Pract Concept 5:29–38 . https://doi.org/10.5826/dpc.0502a04
17. Freeman SR, Greene RE, Kimball AB, Freiman A, Barzilai DA, Muller S, Duke JK, Dellavalle RP (2008) US dermatology residents' satisfaction with training and mentoring: survey results from the 2005 and 2006 Las Vegas Dermatology Seminars. Arch Dermatol 144:896–900 . https://doi.org/10.1001/archderm.144.7.896
18. Freiman A, Barzilai DA, Barankin B, Natsheh A, Shear NH (2005) National appraisal of dermatology residency training: a Canadian study. Arch Dermatol 141:1100–1104 . https://doi.org/10.1001/archderm.141.9.1100
19. Marino ML, Carrera C, Marchetti MA, Marghoob AA (2016) Practice Gaps in Dermatology. Dermatol 34:353–362 . https://doi.org/10.1016/j.det.2016.03.003

20. Roediger HL, Karpicke JD (2006) Test-Enhanced Learning: Taking Memory Tests Improves Long-Term Retention. Psychol Sci 17:249–255 . https://doi.org/10.1111/j.1467-9280.2006.01693.x
21. Kromann CB, Jensen ML, Ringsted C (2009) The effect of testing on skills learning. Med Educ 43:21–27 . https://doi.org/10.1111/j.1365-2923.2008.03245.x
22. Case SM, Swanson DB (2001) Constructing Written Test Questions For the Basic and Clinical Sciences, 3rd ed. National Board of Medical Examiners, Philadelphia
23. Downing SM, Yudkowsky R (2009) Assessment in Health Professions Education, 1 edition. Routledge, New York
24. Schuwirth LWT, van der Vleuten CPM (2003) Written assessment. BMJ 326:643–645
25. Eide MJ, Asgari MM, Fletcher SW, Geller AC, Halpern AC, Shaikh WR, Li L, Alexander GL, Altschuler A, Dusza SW, Marghoob AA, Quigley EA, Weinstock MA, for the INFORMED (INternet course FOR Melanoma Early Detection) Group (2013) Effects on Skills and Practice from a Web-Based Skin Cancer Course for Primary Care Providers. J Am Board Fam Med 26:648–657 . https://doi.org/10.3122/jabfm.2013.06.130108
26. Roads BD, Xu B, Robinson JK, Tanaka JW (2018) The easy-to-hard training advantage with real-world medical images. Cogn Res Princ Implic 3:38 . https://doi.org/10.1186/s41235-018-0131-6
27. Robinson JK, Jain N, Marghoob AA, McGaghie W, MacLean M, Gerami P, Hultgren B, Turrisi R, Mallett K, Martin GJ (2018) A Randomized Trial on the Efficacy of Mastery Learning for Primary Care Provider Melanoma Opportunistic Screening Skills and Practice. J Gen Intern Med 33:855–862 . https://doi.org/10.1007/s11606-018-4311-3
28. Tschandl P, Codella N, Akay BN, Argenziano G, Braun RP, Cabo H, Gutman D, Halpern A, Helba B, Hofmann-Wellenhof R, Lallas A, Lapins J, Longo C, Malvehy J, Marchetti MA, Marghoob A, Menzies S, Oakley A, Paoli J, Puig S, Rinner C, Rosendahl C, Scope A, Sinz C, Soyer HP, Thomas L, Zalaudek I, Kittler H (2019) Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study. Lancet Oncol 20:938–947 . https://doi.org/10.1016/S1470-2045(19)30333-X
29. Keeney S, Hasson F, McKenna HP (2001) A critical review of the Delphi technique as a research methodology for nursing. Int J Nurs Stud 38:195–200
30. Malvehy J, Puig S, Argenziano G, Marghoob AA, Soyer HP (2007) Dermoscopy report: Proposal for standardization. J Am Acad Dermatol 57:84–95 . https://doi.org/10.1016/j.jaad.2006.02.051
31. Harrington E, Clyne B, Wesseling N, Sandhu H, Armstrong L, Bennett H, Fahey T (2017) Diagnosing malignant melanoma in ambulatory care: a systematic review of clinical prediction rules. BMJ Open 7: . https://doi.org/10.1136/bmjopen-2016-014096
32. Turoff M, Linstone HA (2002) The Delphi Method: Techniques and Applications. Addison-Wesley, London
33. A technique for the measurement of attitudes (Book, 1932) [WorldCat.org]. https://www.worldcat.org/title/technique-for-the-measurement-of-attitudes/oclc/812060. Accessed 5 Jul 2019
34. Bloch R, Norman G (2012) Generalizability theory for the perplexed: A practical introduction and guide: AMEE Guide No. 68. Med Teach 34:960–992 . https://doi.org/10.3109/0142159X.2012.703791
35. Haladyna TM (2013) Developing and Validating Test Items, 1 edition. Routledge, New York, NY
36. Jørgensen M, Konge L, Subhi Y (2018) Contrasting groups' standard setting for consequences analysis in validity studies: reporting considerations. Adv Simul 3: . https://doi.org/10.1186/s41077-018-0064-7
37. 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 . https://doi.org/10.1159/000362884
38. 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 thoracoscopic surgery (VATS) lobectomy. Surg Endosc 29:2598–2604 . https://doi.org/10.1007/s00464-014-3975-y
39. Streiner D, Norman G (2008) Divising the items. In: Streiner DL, Norman GR (eds) Health Measurement Scales: a practical guide to their development and use. Oxford University Press, Oxford, pp 18–38
40. Graham B, Regehr G, Wright JG (2003) Delphi as a method to establish consensus for diagnostic criteria. J Clin Epidemiol 56:1150–1156
41. Gruber H (2001) Acquisition of expertise. In: Smelser NJ, Baltes PB (eds) International encyclopedia of the social & behavioral sciences. Elsevier, Amsterdam, pp 5145–5150
42. Patel V, Kaufman D (2001) Cognitive psychology of medical expertise. In: Smelser NJ, Baltes PB (eds) International encyclopedia of the social & behavioral sciences. Elsevier, Amsterdam, pp 9515–9517
43. Hasson F, Keeney S, McKenna H (2000) Research guidelines for the Delphi survey technique. J Adv Nurs 32:1008–1015

44. Jørgensen M, Savran MM, Christakopoulos C, Bek T, Grauslund J, Toft PB, Ziemssen F, Konge L, Sørensen TL, Subhi Y (2019) Development and validation of a multiple-choice questionnaire-based theoretical test in direct ophthalmoscopy. Acta Ophthalmol (Copenh). https://doi.org/10.1111/aos.14065
45. Haenssle HA, Fink C, Schneiderbauer R, Toberer F, Buhl T, Blum A, Kalloo A, Hassen ABH, Thomas L, Enk A, Uhlmann L, Reader study level-I and level-II Groups (2018) Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Ann Oncol Off J Eur Soc Med Oncol 29:1836–1842 . https://doi.org/10.1093/annonc/mdy166
46. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S (2017) Dermatologist-level classification of skin cancer with deep neural networks. Nature 542:115–118 . https://doi.org/10.1038/nature21056
47. Mar VJ, Soyer HP (2018) Artificial intelligence for melanoma diagnosis: how can we deliver on the promise? Ann Oncol 29:1625–1628 . https://doi.org/10.1093/annonc/mdy193
48. Aractingi S, Pellacani G (2019) Computational neural network in melanocytic lesions diagnosis: artificial intelligence to improve diagnosis in dermatology? Eur J Dermatol 29:4–7 . https://doi.org/10.1684/ejd.2019.3538

Figure 1- Content development and validity testing for content through Delphi-like review process.

Figure 2- Plot showing the distribution of items within the various levels of quality: level I (orange), level IB (orange + red lines), II (yellow), III (pink), and IV + remaining items (blue). Level IV and remaining items were either very difficult (DI<0.25), very easy (DI>0.91), or did not possess a sufficient dicriminative ability (DsI<0.1). The 25 items within the yellow, orange and pink areas where included in the final test.

Figure 3 - Box-and-whisker plot showing median scores (black horizontal lines), 95% confidence intervals (grey boxes), range (vertical whiskers) and outlier (circle) for test score results grouped by level of experience.

Figure 4 – Contrasting groups' method: graph shows the normal distribution of novice's (white) and expert's (black) test scores. The point of intersection between the two lines represents the number of correct answers that will result in the maximum number of novices failing the test while as many experts as possible pass the test.

Figure 5 - Decision Test showing the generalizability coefficient as a function of number of test items. The two stapled lines represent number of items needed for an acceptable (Generalizability coefficient > 0.7, n=12) and a good reliability (Generalizability coefficient > 0.8, n=20).

Litterature Review

Expert Interviews

Test Development

Content validation

Test Draft 1
(90 Items)

Review Round 1
+
Language Editing

Average
Quality or Relevance
< 4

Test Draft 2
(60 Items)

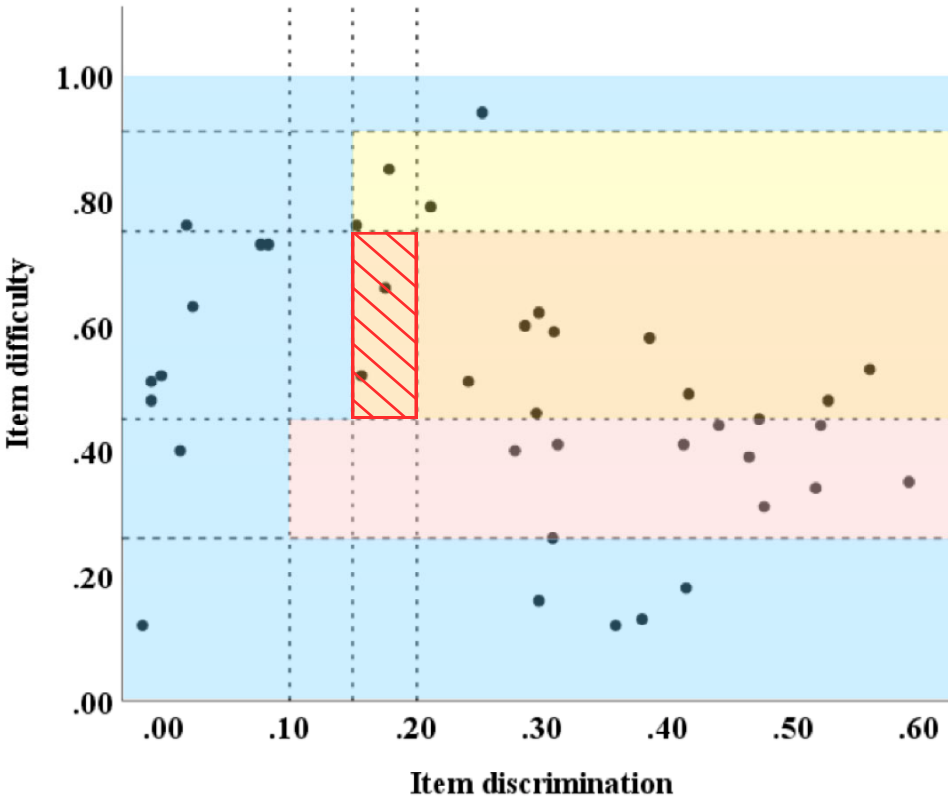
Excluded (30 Items)

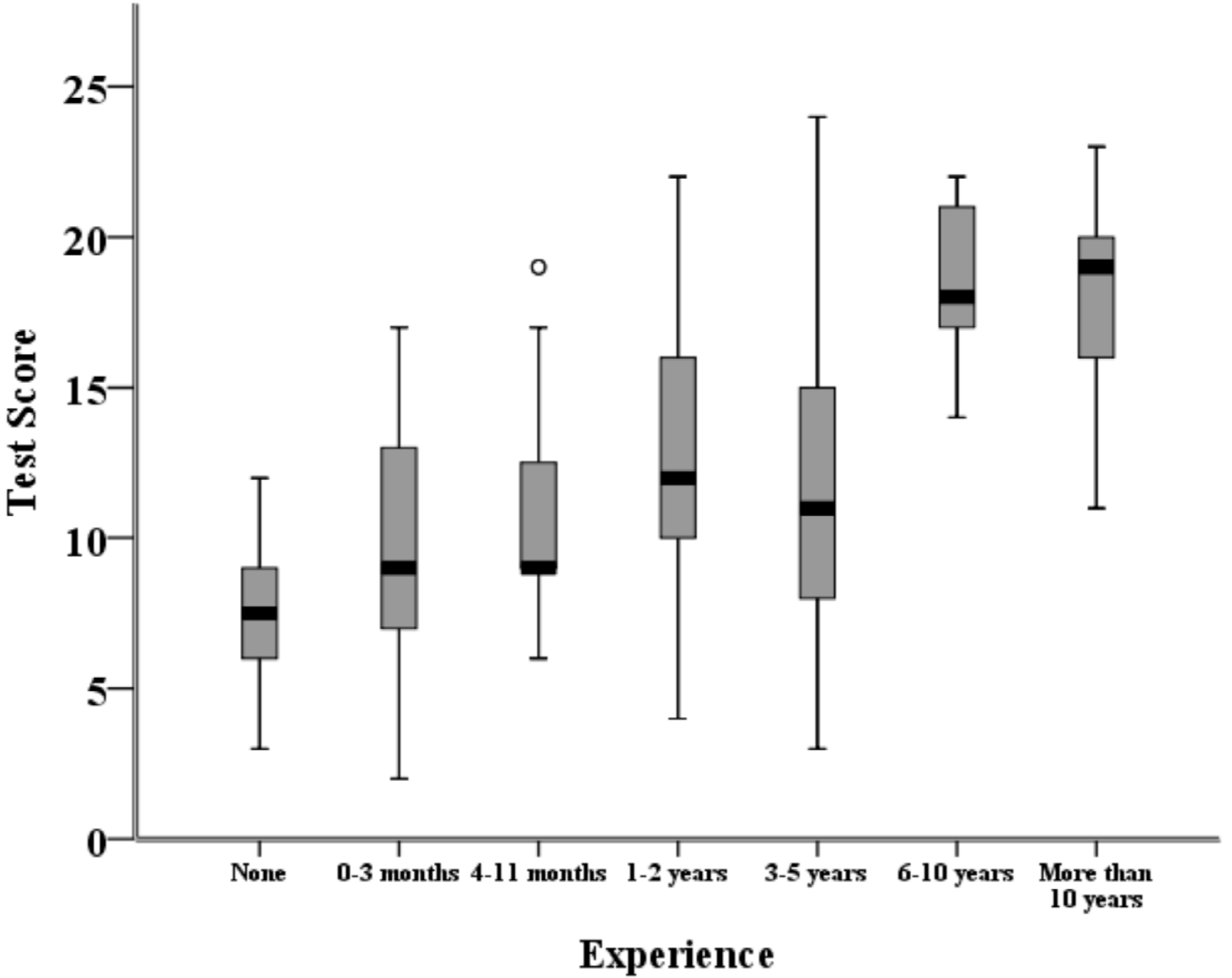
Review Round 2

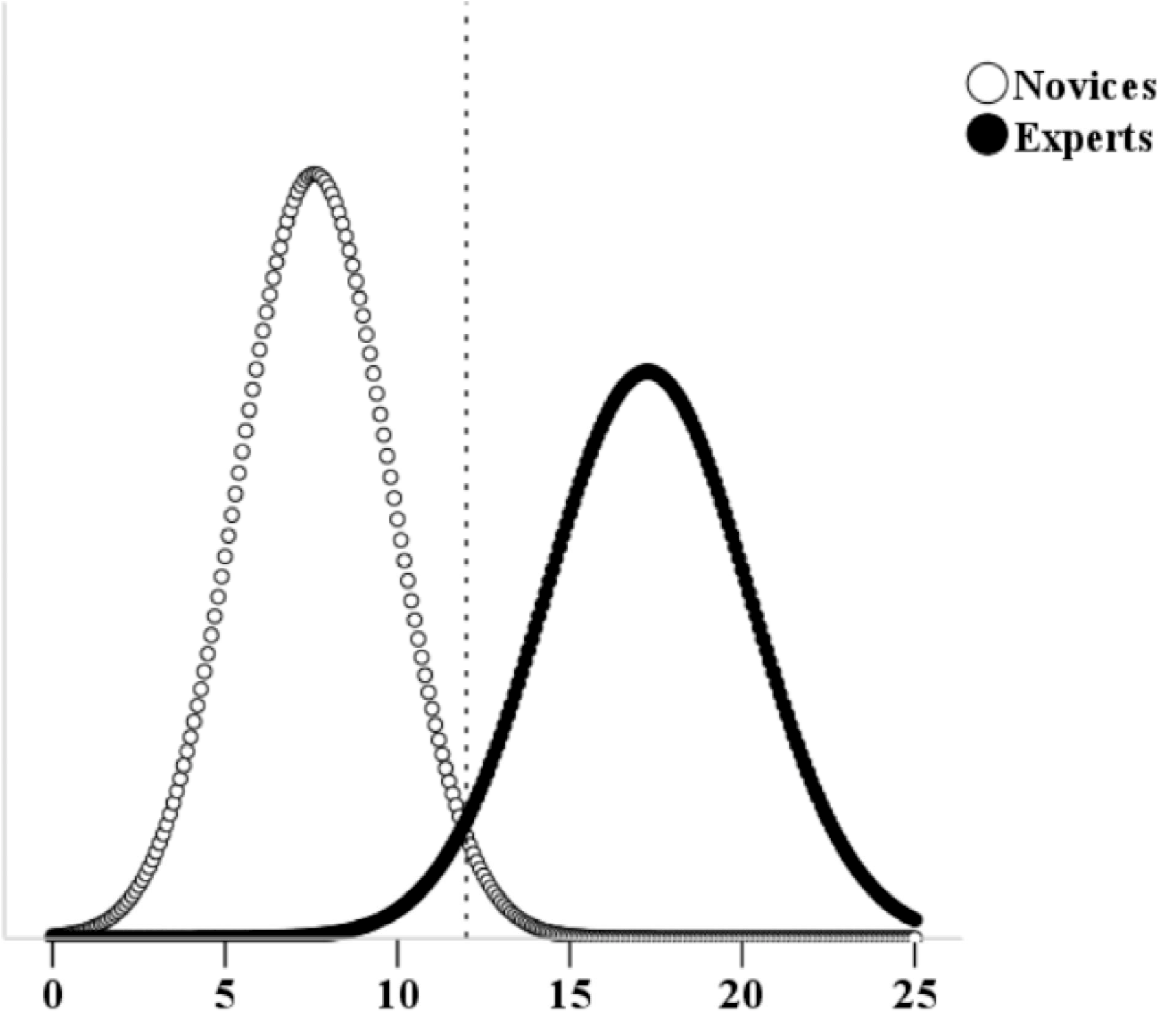
Redundant or Ambiguous

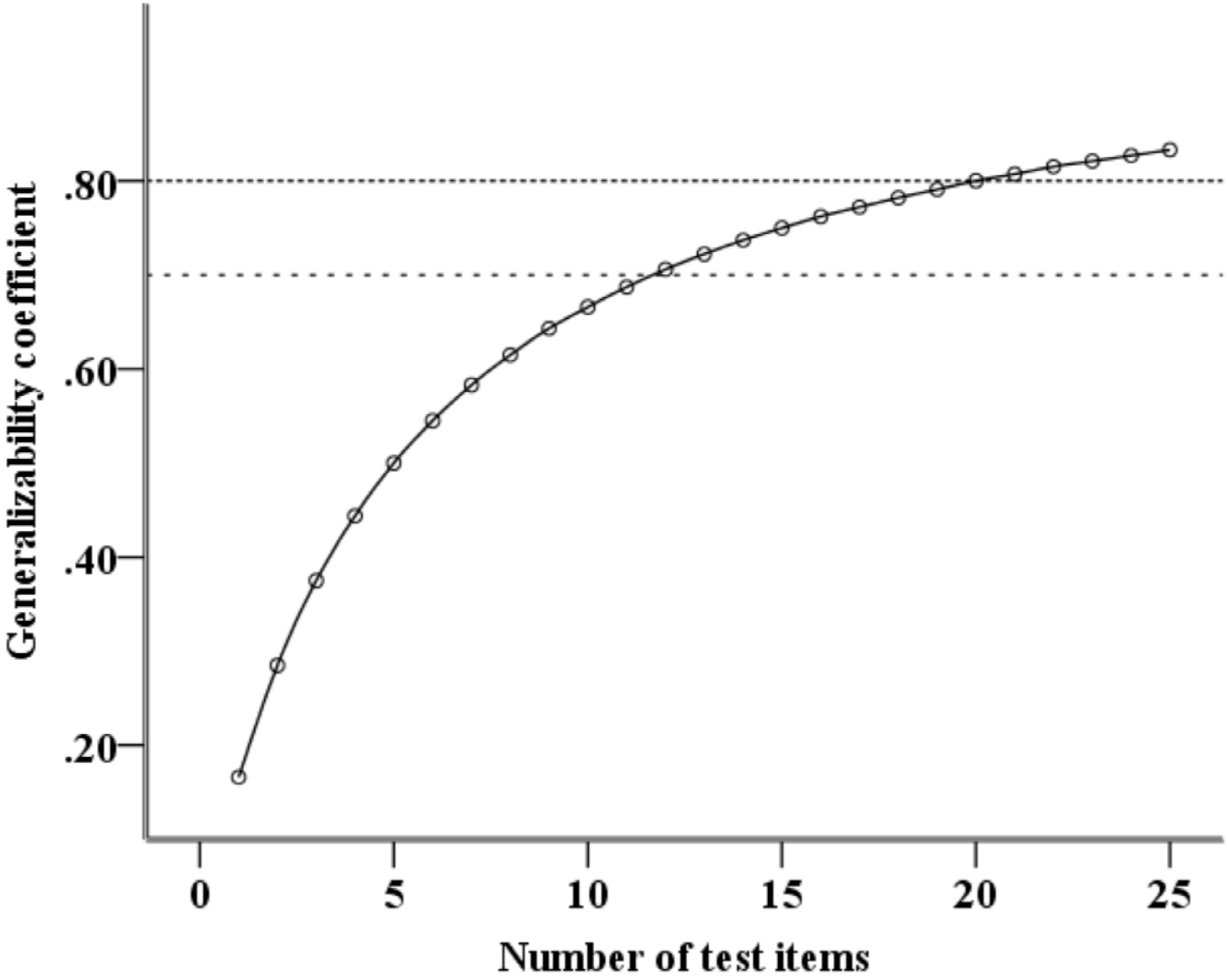
Final Test Draft
(40 Items)

Excluded (20 Items)


1.00
.80
.60
.40
.20
.00
Item difficulty
.00
.10
.20
.30
.40
.50
.60
Item discrimination


Test Score
25
20
15
10
5
0
None
0-3 months
4-11 months
1-2 years
3-5 years
6-10 years
More than 10 years
Experience


Novices
Experts
0
5
10
15
20
25


Generalizability coefficient
.80
.60
.40
.20
0
5
10
15
20
25
Number of test items