Evidence-based medicine in general practice specialty training
Zwolsman, S.E.

Citation for published version (APA):
Zwolsman, S. E. (2012). Evidence-based medicine in general practice specialty training

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

Download date: 30 Nov 2018
Chapter 4

The Dutch Berlin Questionnaire: translation and validation

Sandra E. Zwolsman, Margreet Wieringa – de Waard, Lotty Hooft, Nynke van Dijk

A short version of this article has been published in J Clin Epidem 2011;64:928-30
An adapted Dutch translation of this articles has been published in NTvEBP 2011;9:8-11
Abstract

**Objective**  To determine the reliability and validity of the Dutch version of the Berlin questionnaire.

**Study Design and Setting**  The Berlin questionnaire measures EBM knowledge and skills. The two sets of the Berlin questionnaire were translated into Dutch and filled out by 140 general practice trainees, and by 37 course participants and seven tutors of Dutch Cochrane EBM courses. For both sets reliability and validity were assessed.

**Results**  Face validity assessment resulted in positive responses regarding the translation of the questionnaire. Reliability measures of the translated questionnaire are acceptable: a few questions had either very low or high percentages of correct answers. No trainees scored the highest or lowest possible total score on either set. Item-total reliability of some questions and the internal consistency of the test are lower than expected. Validity is high. Scores of the course participants increased after the course and course participants scored lower on both sets than the course tutors did.

**Conclusion**  The Dutch version of the Berlin questionnaire reliably and validly measures EBM knowledge and skills. Scores on Set 1 are higher than scores on Set 2, which makes the outcomes of both sets incomparable without correcting for score differences.
CHAPTER 4

Introduction

Evidence-based medicine (EBM) can be defined as “the conscientious, explicit and judicious use of current best evidence, in combination with the physician’s clinical expertise, patient preferences and likely actions of the patient, in making decisions about the care of individual patients”\textsuperscript{1,5} In the Netherlands, EBM is part of the curriculum of the Master of Medical Science program and of various specialty training programs.

For the evaluation of the effectiveness of EBM education a validated instrument is needed.\textsuperscript{6,7} One of the few validated instruments for measuring knowledge of and skills related to EBM is the Berlin questionnaire.\textsuperscript{6} The Berlin questionnaire has been used in multiple studies\textsuperscript{8-10} and is written in English.\textsuperscript{6} Although most Dutch students can adequately read and write the English language, a Dutch version of the Berlin questionnaire would ensure that all questions and answers are easily interpretable and entirely understandable.\textsuperscript{11}

The aim of this study is to examine the reliability and validity of a Dutch version of the Berlin questionnaire in assessing EBM knowledge and skills. To do so, we assessed face validity, the difficulty of the questions, frequency of endorsement, range and distribution of scores, internal consistency, item-total correlation, responsiveness to change and discriminant validity.\textsuperscript{11-14}

Methods

Berlin questionnaire

The Berlin questionnaire was designed by Fritsche et al.\textsuperscript{6} in order to measure the basic knowledge and skills required for practising EBM. The questionnaire consists of two sets (Set 1 and Set 2) of 15 multiple-choice questions with 5-answer options of similar content and can therefore be
used for before and after measurements.\textsuperscript{6} In the original study, the questionnaire was shown to be reliable and valid.\textsuperscript{7}

\textit{Translation}

Both sets of the questionnaire were translated into Dutch using forward-backward translation.\textsuperscript{12} First, a preliminary translation from English into Dutch was done independently by two Dutch natives both fluent in English. Differences between the two Dutch versions were discussed, resulting in the re-writing in one Dutch translation of both sets of the questionnaire. This version was then retranslated into English by a professional translator who was blinded to the original version. Differences between these versions and the original English version were discussed by the first two translators and changes were made if necessary. Inconsistencies and apparent errors in the original questionnaire led to corrections in the Dutch translation. This resulted in one final Dutch translation of both sets of the questionnaire.

\textit{Participants}

To determine the reliability and validity of the questionnaire, three different groups of subjects took part in our study.

In order to determine how the Dutch translation was perceived (face validity),\textsuperscript{11} the translated questionnaire was presented to a mixed panel comprising GPs and scientists with different levels of research experience and EBM expertise.

For reliability assessment, 140 first year GP trainees from the three-year competency based\textsuperscript{15} GP specialty training programmes at the Academic Medical Center, University of Amsterdam, the University of Groningen and Leiden University filled out the questionnaire under exam-conditions.\textsuperscript{5} They completed the questionnaire in the first month of their specialty training.
CHAPTER 4

The Amsterdam trainees completed Set 1 of the questionnaire, the trainees from Groningen and Leiden Set 2.

To assess the validity of the questionnaire (the responsiveness to changes in knowledge and skills within subjects and difference in knowledge and skills between groups of subjects) 37 participants and seven tutors of two analogous EBM courses given by the Dutch Cochrane Centre were included. These three-day courses on ‘EBM in clinical practice’ provide the participants with the knowledge and skills needed to apply the results from evidence in practice. The course participants were professionals from various healthcare settings and trainees from various clinical specialty training programs. The courses were given in January 2009 (group A) and March 2009 (group B). To avoid effects of differences in the difficulty of both sets, half of all the participants randomly filled in Set 1 before and Set 2 after the course, the other half filled in Set 2 before and Set 1 after the course. One week prior to the course, Group A received the questionnaire (either Set 1 or 2) at their home address, and were asked to complete this without using books or other tools. After the course, respondents to the first questionnaire were asked to complete the other set at the last session of the course. The participants of Group B completed the first questionnaire (either Set 1 or Set 2) at the first day of the course before discussion of any relevant content, and the other set during the last session. Three tutors filled out Set 1 and four tutors Set 2 of the questionnaire.

The directors of the training programs and courses approved of the study and the administration of the questionnaires. The participants completed the questionnaire voluntarily and all gave written informed consent. Anonymity of the participants was ensured by coding the questionnaires with a study number, of which the key was kept separate from the questionnaire data in a locked cabinet.
Questionnaire Characteristics

We assessed the face validity, and the reliability and validity of the questionnaire. The reliability of the questionnaire was determined to assess whether the questionnaire measures EBM knowledge and skills accurately. Reliability was determined by assessing the difficulty and scoring of the individual questions (proportion of correct answers) and of the questionnaire as a whole (range of total scores and score-distribution), the answer options most often chosen (“frequency of endorsement”), the coherence of all the questions (internal consistency or alpha) and the correlation of the individual questions to the test as a whole (item-total correlation). The validity of the questionnaire was determined to assess whether the questionnaire does indeed measure EBM knowledge: the ability to measure increase in knowledge and skills after a content-relevant course (responsiveness) and the ability to discriminate between groups of different levels of EBM knowledge and skills (known-group comparison).

Face validity

We asked the panel to critically assess the contents and lay-out of the questionnaire, the language and terminology used, and to judge whether the cases presented in the questionnaire suited the Dutch medical setting. To systematically gain the information needed, a one-page checklist with four closed-ended questions and a free text box for comments regarding content and clinical relevance of the translated questionnaire was developed. The comments of the panel members were reviewed by the translators and, where necessary, the Dutch version of the questionnaire was adjusted.
Reliability

For the assessment of reliability the data from the 140 GP-trainees was used.

The difficulty of the questions was determined by calculating the number of participants with missing answers and the proportion of participants with correct answers for each individual question. We considered a range of 10 to 90% correct answers acceptable. Furthermore the ‘frequency of endorsement’, which is the proportion of subjects that chose each (in)correct answer-option, was calculated to assess their value. If an incorrect answer-option was chosen by less than 5% of the participants, this answer-option was considered less useful, since it indicates that certain answer options are considered unrealistic by the participants and are therefore not chosen. The choice for the correct answer is then based on the fact that other answer-options are clearly unlikely, instead of knowledge on the correct answer.

Additionally, the range and distribution of total scores (i.e. skew, floor and ceiling effects) were verified. We considered difficulty satisfactory if less than 15% of participants scored the lowest (0) or highest (15) possible total score, since otherwise no improvement or deterioration of knowledge and skills can be assessed.

To evaluate the correlation of the individual questions, and thereby determine the homogeneity and unidimensionality of each set of the questionnaire, the internal consistency was calculated using KR-20s alpha for dichotomous data. An alpha of >0.7 is considered satisfactory. The item-total correlation was calculated to assess the correlation between the individual items and the test as a whole. A value of >0.2 for each question was considered satisfactory.

All the above-mentioned tests were performed separately on both sets of the questionnaire. The mean scores from both sets were compared to assess differences in difficulty of both sets and thereby the usefulness of
the questionnaire in comparing differences between groups or changes in knowledge and skills.

Validity

The assessment of validity of the questionnaire was performed on the questionnaires collected from the participants (n = 37) and tutors (n = 7) at the EBM-course. The responsiveness was determined by the change in mean score after attending a theoretical EBM course. This was assessed by comparing the scores of the participants before and after the course.

The discriminant validity was determined by known-group comparison. The purpose of the known-group comparison was to assess whether the Berlin questionnaire discriminated between subgroups that are expected to differ in EBM-knowledge and skills. For this purpose the results of EBM experts (course tutors) were compared with the scores of the non-experts (course participants) before the start of the course. Tutors were expected to score significantly higher than participants.

Analysis

Characteristics of the participants and overall scores of the questionnaires were described by proportions for categorical data, means and standard deviations for normally distributed continuous data and medians and quartiles for non-normally distributed continuous data. We analyzed the questionnaires as described in the original study. Total scores on the questionnaire were expressed as means and standard deviations. Differences in scores among groups tested with Set 1 and Set 2 or with different levels of expertise were compared using the Student’s t-test. To calculate pre- and post-course differences a paired t-test was used. An increase of the mean score at least 2.0 points was expected, in accordance with the study of Fritsche et al. A power analysis showed that 21 participants were needed to find a difference of 2 points, with an alpha of
0.05, power of 80%, and a standard deviation of 2.2. Correction for the sequence of administration (either Set 1 or Set 2 before the course) was done in case differences between the mean scores (difficulty) of both sets of the questionnaire were found, using ANCOVA. For the assessment of the responsiveness of the Berlin questionnaire the results of Set 1 and Set 2 were combined. A p-value of < 0.05 was considered statistically significant. Analyses were done using SPSS version 16.0.

Results

Participants

In total 140 trainees, 37 course participants and seven tutors filled in questionnaires (Table 1):
- GP trainees completed either Set 1 (n=73) or Set 2 (n=67) of the Dutch version of the Berlin questionnaire.
- Out of 83 course participants (67 from group A and 16 from group B), 37 (45%) completed both sets of the questionnaire. Reasons for not handing in the questionnaire were absence (n=5), not having received the initial questionnaire by post (n=7), no time to fill in the questionnaire (n=6), no possibility to match (n=11); missing signatures (n=2); reasons unknown (n=15).
- Out of ten course tutors, seven completed either Set 1 (n = 3) or Set 2 (n = 4) of the Dutch version of the Berlin questionnaire. The three other tutors were involved in the design of this study and therefore did not complete the questionnaire.

Face validity

Four panel members were positive about the Dutch version of the Berlin questionnaire and thought it to be helpful in measuring EBM knowledge.
and skills. One member stated that the questions were hard for the average GP to answer. Also, all except one agreed that the translated sets of the questionnaire suited the Dutch clinical setting and were linguistically correct. The person who did not agree stated that the original questionnaire should have been translated more freely. Overall, the panel members agreed that the Dutch version of the Berlin questionnaire was relevant and useful. Their remarks did not result in any changes in the questionnaire.

Table 1. Characteristics of the study participants.

<table>
<thead>
<tr>
<th></th>
<th>GP trainees</th>
<th>Course participants</th>
<th>Course tutors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number</td>
<td>140</td>
<td>37</td>
<td>7</td>
</tr>
<tr>
<td>Age (mean ± SD)</td>
<td>29.9 ± 6.8</td>
<td>37.3 ± 10.2</td>
<td>37.7 ± 7.5</td>
</tr>
<tr>
<td>Years since graduation</td>
<td>2 (1.25-3)</td>
<td>5.5 (3-14)</td>
<td>14 (8.5-20.5)</td>
</tr>
<tr>
<td>Women</td>
<td>71.4%</td>
<td>55.6%</td>
<td>71.4%</td>
</tr>
<tr>
<td>PhD</td>
<td>-</td>
<td>16.7%</td>
<td>71.4%</td>
</tr>
<tr>
<td>Profession</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Health care</td>
<td>100%</td>
<td>75.7%</td>
<td>28.6%</td>
</tr>
<tr>
<td>• Scientific</td>
<td>-</td>
<td>13.5%</td>
<td>71.4%</td>
</tr>
<tr>
<td>• Other</td>
<td>-</td>
<td>10.8%</td>
<td>-</td>
</tr>
</tbody>
</table>

Reliability

To assess reliability, the questionnaire outcomes of the 140 GP trainees were used.

- Difficulty of the questions

One question from Set 1 had a specifically high number of missing answers (26%). This question regards a statistical calculation on likelihood ratio from
a two-by-two table. The number of missing answers for the other questions ranged from 0% - 11.3% (Set 1) and 0% - 15.9% (Set 2).

A question was considered of adequate difficulty if between 10-90% of the participants chose the correct answer. Set 1 contains one question that 90.4% of respondents answered correctly and therefore can be considered too easy, not discriminating poor from good respondents. All other questions were answered correctly by 14.5% - 85.9% of participants. Set 2 contains two questions with <10% correct answers (4.5% and 9.1%) and one question that was answered correctly by 96.8%. These questions show less discriminative value between poor and good respondents. All other questions were answered correctly by 21.9% - 81.5% of participants. Set 1 contained 4 questions and Set 2 contained 7 questions of which one of the incorrect answer options was chosen more frequently than the correct answer-option.

An incorrect answer option that was chosen by < 5% of the participants is considered less useful. This was the case for all but 2 questions in both Set 1 and Set 2, although both sets contained only 2 questions with answer-options that were not chosen by any participant. This indicates that these questions can be considered less difficult since the choice for the correct answer is based on which answer-options are clearly not realistic, instead of knowledge on the correct answer.

- Difficulty of the whole questionnaire
The distribution of total scores was explored to see whether the respondents find the questionnaire too easy or too difficult (more than 15% of participant scoring the lowest or highest possible score). None of the trainees scored the highest or lowest possible total score on either Set 1 or Set 2. (Figure 1) The total scores were normally distributed. The mean total score on Set 1 was 7.6±2.2 (maximum score 15 points). The mean total score on Set 2 was 6.0±2.3 (maximum score 15 points). This difference in scores was statistically significant (mean differences 1.5 points; 95% CI 0.8 – 2.3 points).

- Internal consistency
The questionnaire is considered internally consistent if alpha reaches >0.70. Set 1 has an alpha of 0.45. Set 2 has an alpha of 0.50.

- **Item total reliability**

The item total reliability is considered sufficient if it reaches >0.2. Three questions on each set showed a correlation coefficient that falls below the 0.2 level (Set 1 0.07 – 0.17; Set 2 0.15 – 0.18).

**Validity**

To assess validity, the questionnaire outcomes of course participants (n = 37) and from the course tutors (n = 7) were used.

- **Responsiveness**

Mean scores of course participants before the course were 7.4±2.5 and after the course 9.0±2.2 (Figure 2). Scores increased by 1.5 (95% CI 0.5 – 2.5 points), irrespective of the sequence of administration of the two sets of the questionnaire. Sequence of administration of the sets of the questionnaire predicted a score difference (p=0.011), which means that the height of the scores obtained depends on which set of the questionnaire was assessed first.

- **Known groups**

Course participants scored lower than experts, namely 7.4±2.5 versus a mean of 12.4±1.0 points scored by the course tutors; mean difference 5.0 points (95%CI 3.9 – 6.1 points).
Figure 1. Scores of GP trainees (n = 140) on Set 1 and Set 2 of Dutch version of Berlin questionnaire

Figure 2. Before and after scores of course participants (n = 37) on Dutch version of Berlin questionnaire
Discussion

We translated the Berlin Questionnaire, developed by Fritsche et al. to measure EBM knowledge and skills\(^6\) into Dutch by using the forward-backward method, and assessed the reliability and validity of this Dutch version among GP-trainees and among course participants and tutors of an EBM-course. We found that, like the original English version, the Dutch instrument was suitable for measuring EBM knowledge and skills. Face validity analysis revealed that the Dutch translation was accepted by both researchers and GPs, and GPs with research experience. Also the other reliability and validity measures (difficulty, responsiveness, known-group comparison) were adequate. EBM expert score significantly higher than EBM “layman”, and teaching EBM leads to a significant increase in scoring on the questionnaires.

Reliability of the instrument is adequate with only a few questions scoring low on discriminative value or inadequate answer options. One outcome of the reliability analysis needs specific attention. Looking at the mean scores of GP trainees, the total mean scores of Set 1 do not correspond with the mean scores of Set 2. Analyses showed that mean scores of Set 1 are 1.5 points higher, regardless the time of administration (before/after course). In this study this problem was solved by the sequence of administration: 50% of the respondents filled in Set 1 before and Set 2 after the course and the other 50% Set 2 before and Set 1 after the course. Another way of solving this problem is using ANCOVA analysis using the sequence of administration as the grouping variable. The original English questionnaire also shows a difference of 0.5 to 0.8 points in mean scores when comparing Sets 1 and 2. However, no further attention was paid to this by Fritsche et al.\(^6\) As shown in the reliability analysis the explanation for the difference in difficulty of the sets could be that Set 2 contains more difficult questions (scored correct by less than 10% of participants), and more questions have incorrect answer options that seem correct to the majority of participants. Another explanation for the
difference between the scores on Set 1 and Set 2 could be that the GP-trainees from the different universities actually differ in EBM knowledge and skills. Since the content of the bachelor and master of science program of Medicine is determined by the government in the Netherlands and described in the “Framework for Undergraduate Medical Education in The Netherlands”, this seems unlikely. Furthermore, a similar difference in scores (mean difference 1.6; 95% CI -0.06 – 3.17; data not presented) is observed between the course participants filling out the different sets before the course. Since they were randomly assigned to either Set 1 or Set 2 a structural difference in difficulty between the sets is probably present.

Scale and item characteristics in our study were acceptable. The scale difficulty is suitable for the population in this study; none of the participants scored either the highest or the lowest possible score. Item difficulty is good with only four questions being relatively easy or difficult on a total of 30 questions. The high number of missings on one question was probably caused by the calculations needed to answer the question.

In the original study of Fritsche et al. internal consistency was higher (0.75 for Set 1 and 0.82 for Set 2) than in our study. The alphas we found in our study are, with a value of 0.45 and 0.5, rather low. This could be an indication that the two sets of the questionnaire are not internally consistent. The alpha may, however, not be fully appropriate to assess reliability of the Berlin questionnaire. Alphas can only be used if the items are linked to one underlying characteristic. As the non-homogenous individual questions with their heterogeneous answer options do not contribute linearly to an overall characteristic, but are independent in measuring the application of a concept, one could discuss whether an alpha should be used. Furthermore, the group in which the alpha is calculated consists of people with comparable levels of knowledge, whereas all trainees have a Masters degree in Medical Science. The homogeneity of the course participants could also contribute to a low alpha. The sample used by Fritsche et al. might have been more heterogeneous. There is no reason to assume that the lower alphas found
in our study are a result of the translation of the questionnaire or reduce the usefulness of the questionnaire.

Validity of the Dutch translation of the Berlin questionnaire is good. The questionnaire shows good responsiveness (increased knowledge after a course) and differentiates between known groups (course participants versus tutors). There are some limitations to the methods we used for assessing the validity of the Dutch version of the Berlin questionnaire. The questionnaire was only assessed by a small group of course tutors. Since their scores did already differ significantly from scores of course participants, the questionnaire seems to discriminate adequately between groups with different levels of expertise.

The response rate of the course participants of the group A was poor. One explanation for the low response rate could be that group A received the ‘before’ set of questions at home. Not all course participants received the questionnaire and some had no time to fill it in before attending the course. Therefore, logistical factors (time, post) are considered the most important reasons for not filling in the questionnaire, which probably does not influence the results. It could also be that participants that considered the questionnaire relatively easy were more likely to respond than those that considered the questionnaire more difficult as they would need more time. This would probably have led to a relatively high number of ‘good’ participants filling in the questionnaire, probably reducing the difference between the before and after course scores, and the difference between participants and tutors, resulting in an underestimation of the true differences. Furthermore the course participants of group A could have used books or other material when filling out the first questionnaire. This would also have led to an underestimation of responsiveness and discrimination between participants and tutors.

In an ideal situation the results on the Berlin questionnaire would have been compared with a ‘gold standard’ test on knowledge and skills regarding EBM. Such an instrument, assessing the ‘true’ knowledge and skills regarding EBM is currently not available. Only the Fresno-test has
also been validated for assessing EBM-knowledge and skills. This test focuses on specific skills needed for evidence-based practice,\textsuperscript{20} such as formulating search questions and performing searches. The Berlin questionnaire focuses on knowledge and skills needed to apply the results of studies in practice.\textsuperscript{6} Because of the different goals of this questionnaire we chose not to assess concurrent validity.

Conclusion

The Dutch translation of the Berlin questionnaire is a reliable and valid measure of EBM knowledge and skills and can be used in assessing trainees of Dutch medical specialty programs. However, the sequence of administration of the two sets of questions should be handled with care because the scores on Set 1 are higher than the scores on Set 2 and are therefore not comparable.