Measuring psychosocial consequences of false-positive screening results - breast cancer as an example
Brodersen, John

Publication date: 2006

Document Version
Early version, also known as pre-print

Citation for published version (APA):
Measuring psychosocial consequences of false-positive screening results – breast cancer as an example

PhD thesis
John Brodersen, MD, GP

Supervisors:
Marjukka Mäkelä, MD, PhD, MSc, professor
Hanne Thorsen, MD, PhD, Senior researcher
Department of General Practice, University of Copenhagen
Submitted March 2006
Accepted for public defence August 2006
PhD thesis

Measuring psychosocial consequences of false-positive screening results – breast cancer as an example

John Brodersen
MD, GP

Faculty of Health Sciences
University of Copenhagen

Copenhagen 2006
The PhD thesis was accepted by the Faculty of Health Sciences of the University of Copenhagen.

Correspondence:
John Brodersen, MD, GP
Department of General Practice
Institute of Public Health
Centre for Health and Society
University of Copenhagen
Øster Farimagsgade 5, 24Q, Postbox 2099
DK-1014 Copenhagen
Denmark
e-mail: j.brodersen@gpmed.ku.dk

Printed in Denmark
By Månedsskrift for Praktisk Lægegerning
Copenhagen
ISBN: 87-88638-36-7
8.2.1. Translation and adaptation ................................................................. 49
8.2.2. Assessment of content validity .......................................................... 50
8.3. Justification of conclusion ...................................................................... 51
8.4. Contribution to the current knowledge ..................................................... 51
9. Article 3 ........................................................................................................ 52
10. Discussion of article 3 .................................................................................. 75
10.1. Major findings ......................................................................................... 75
10.2. Assessment of methods ........................................................................... 75
10.2.1. Data collection .................................................................................... 75
10.2.2. Statistical methods in general ............................................................... 76
10.2.2.1. Sleep items ....................................................................................... 76
10.2.2.2. Response categories and threshold order ........................................... 76
10.2.2.3. Differential item functioning ............................................................ 76
10.2.2.4. Combining Item Response Theory and Classical Test Theory .............. 77
10.2.2.5. Test of concurrent validity ................................................................. 77
10.3. Justification of conclusion ....................................................................... 78
10.4. Contribution to the current knowledge .................................................... 78
11. Article 4 ....................................................................................................... 80
12. Discussion of article 4 .................................................................................. 99
12.1. Major findings ......................................................................................... 99
12.2. Assessment of methods ......................................................................... 99
12.2.1. Assessment of content validity ............................................................. 99
12.2.2. Data collection .................................................................................... 100
12.2.3. Statistical methods in general ............................................................... 101
12.2.3.1. Order of response categories ............................................................. 101
12.3. Justification of conclusion ..................................................................... 101
12.4. Contribution to the current knowledge ................................................... 102
13. The process of the PhD study .................................................................... 103
14. Conclusions and implications for practice and research .............................. 104
14.1. Conclusion ............................................................................................. 104
14.2. Implications for practice and research ..................................................... 105
14.2.1. Implications for practice ..................................................................... 105
14.2.2. Implications for research ................................................................. 105
15. Reference list ............................................................................................. 107
16. Abbreviations and professional terms ........................................................... 116
17. Appendix I - the Psychological Consequences Questionnaire (PCQ) .......... 120
18. Appendix II - overall search strategy .......................................................... 122
19. Appendix III - translation report ................................................................. 123
20. Appendix IV - the PCQ-DK33 (Danish version) ........................................... 136
21. Appendix V - the PCQ-DK33 (English version) ......................................... 141
22. Appendix VI – part 2 of the measure (Danish version) ............................... 146
23. Appendix VII – part 2 of the measure (English version) ............................ 151
1. Preface and acknowledgements

The structure of the thesis

According to the “Rules and regulations, a guide to the PhD study” from the Faculty of Health Science, University of Copenhagen a PhD thesis encompassing more than one article should be accompanied by an account of the articles’ publishing status. The account can be structured as follows:

A. a brief, general presentation of the research problem dealt with by the included articles
B. a brief presentation of the results achieved with an assessment of the methods used and a critical review of the conclusions that can be drawn from the results
C. a comparison with and assessment of published results of other researchers to the extent that they are relevant to the author’s analysis of the research problem
D. a summarising conclusion

This PhD thesis encompasses four scientific papers. The structure of the thesis follows the structure suggested by the University, so that point A covers the introduction, background and aims of the thesis (pages 5 to 25). The introduction is followed by the four scientific articles (pages 26 to 102). Each article is followed by a discussion of the current study. This discussion is structured so that it encompasses the following: major findings, assessment of methods, justification of conclusion and contribution to current knowledge in fulfilment of points B and C. The conclusion of the thesis (page 104) covers point D.

Financial support

This thesis is based on studies carried out during an appointment with the Department of General Practice, Institute of Public Health, University of Copenhagen from 2000 to 2006 funded by external research grants. Therefore, I would like to thank all of the contributors (alphabetically):

- AstraZeneca's rejselegat for almen medicin
- Fonden vedrørende finansiering af forskning i almen praksis og sundhedsvæsenet i øvrigt
- Fridriksons mindelegat
- Margot Fribergs fond
- Praktiserende Lægers Uddannelses- og udviklingsfond
- Det Psykosociale Forskningsudvalg, Kæftens Bekæmpelse
- Sygekassernes Helsefond
- Tvergaard Fonden

The origins of this project

As a medical student and a doctor I was taught, above all else, to do no harm. Therefore, it was an eye opener for me when Inga Marie Lunde, a general practitioner, demonstrated at the theoretical course of the specialist education of general practice how medical screening has the potential to be as harmful as it is beneficial. Jack Hoffman, head of the Breast Surgical Clinic at Hørsholm Hospital where I worked as a trainee, was the second person to spark my interest in this area. Our discussions about the reactions of women to false positive diagnostics in the outpatient treatment area, where women with breast tumours were examined were invaluable. The third person I wish to thank is John Sahl Andersen, assistant professor and general practitioner, who encouraged me to begin this project. I am very thankful for the support John gave me in the beginning of my project and for his role as my supervisor during my first years as a PhD student.
Supervisors and collaborators on the project

It was simply by chance that I meet Hanne Thorsen, senior researcher, at the Department of General Practice, University of Copenhagen. Nevertheless, Hanne’s great enthusiasm, knowledge and generosity in the research area of questionnaires and measurement has been invaluable to me in planning this project and has brought the standards of the methods used to highest attainable level. Because of Hanne’s mentorship, this project is significantly more academic, relevant and exciting. I will be forever grateful for her introduction to the world of psychometrics.

My supervisor professor Marjukka Mäkelä taught me how to structure a report and how to organise the writing process. Because of Marjukka’s in depth knowledge of breast cancer screening and teaching ability, writing the account of this thesis became a pleasure. I am very thankful for Marjukka’s help and for accepting the role as my supervisor for the last period of my PhD project.

I would like to thank my clinical supervisor, head of the Breast Cancer Screening Clinic in Copenhagen, Ilse Vejborg, for her critical comments and for assisting me in getting access to the women screened. In relation to this, I also wish to thank the staff at the two screening clinics (Copenhagen and Funen) for their collaboration and help in recruiting the women screened. A special thanks is given to all of the secretaries at the two screening clinics for their help identifying relevant women for interviews, asking women for participation in the questionnaire survey and mailing questionnaires to women in the control group.

I am in great debt to the late professor Jill Cockburn for her strong support and encouragement in this project and for her inspiration and co-authorship in the first paper in this thesis.

Last but not least, I would like to express my sincere gratitude to the more than one thousand women that have either participated as interviewees or respondents in this project. Without their contribution I would not have been able to conduct this thesis.

Other supportive individuals and environments

Being a part of the research environment of general practice at the University of Copenhagen has been essential for me as a PhD student. For their contribution to making this environment the great resource it is, I would like to thank all of my present and former colleagues at the Department and Research Unit of General Practice. In addition, I would like to give a special thanks to Anja Drikkjær, Lise Hoyer and Lone Merio for their secretarial assistance.

I wish to acknowledge the support from fellow members of the Risk Group and the Screening Group, both former interest-groups attached to the Danish College of General Practitioners. I also wish to thank all the fellow members of the Nordic Risk Group, which is a group of academically active Nordic general practitioners sharing the vision “to promote general practice which is salutogenic, empowering and sustainable” and the fellow members of the Matilda Bay Club, which is a group of academics using Rasch analysis. In addition, I would like to give Svend Kreiner, assistant professor, a special thanks for our many and fruitful talks about Rasch analysis.

Finally, I wish to thank my friends, mother, parents-in-law and family for their support, tolerance, help and love. To my children, Franciska and Markus - I thank you most of all for inspiring me with your happiness and joy of life. To my wife Annette - without your love, support, courage and hard work, especially in relation to the two periods with no financial support, I would not have made it - thank you!

John Brodersen
Copenhagen, August 2006
2. **Summary**

**Objectives**

The overall purpose of this PhD study was to identify, adapt and validate an instrument to measure the psychosocial consequences of false-positive screening mammography (hereafter referred to as false positives) in Denmark. The thesis encompasses four scientific articles covering the aims:

- to review quantitative studies on women’s experiences of short-term and long-term consequences of false positives and to identify and assess the adequacy of the most frequently used instruments for measuring these consequences
- to translate and adapt into Danish the negative part of the Psychological Consequences Questionnaire and to assess the content validity of the Danish version of the instrument in the setting of abnormal screening mammography later confirmed to be false-positive
- to statistically validate a new condition-specific instrument measuring psychosocial consequences of abnormal screening mammography
- to develop and validate a new condition-specific instrument measuring long-term psychosocial consequences of false positives

**Material and method**

A systematic literature review was conducted identifying papers reporting quantitative studies on the consequences of false positives using the MEDLINE, CINAHL, EMBASE and PsycInfo databases. Articles citing the development and psychometric properties of the most frequently used measures were also retrieved. Finally, the review focused on studies that had used at least one of the most frequently used measures.

The translation and adaptation into Danish of the most adequate measure included a bilingual panel and a lay panel. Content validity was ensured by conducting six focus group interviews including 34 women who had previously received an abnormal screening mammography that was later confirmed to be a false positive. The acceptability of the instrument was ensured by conducting fifteen individual telephone interviews with women who had had an invitation to attend breast cancer screening and with women who had never been invited and screened.

The draft version of the new instrument with high content validity for women having false positives was completed by women who had received an abnormal, a false-positive and a normal screening mammography. Item Response Theories (the Rasch model) and Classical Test Theories were used to analyse data. Construct validity, concurrent validity, known group validity, dimensionality, additivity, objectivity, internal consistency and reliability were established.
Results

Twenty-three quantitative studies on women’s experiences of short-term and long-term consequences of false positives were identified. The most commonly used measures were the General Health Questionnaire (GHQ), the Hospital Anxiety and Depression Scale (HADS), the Psychological Consequences Questionnaire (PCQ) and the State-Trait Anxiety Inventory (STAI). One or more of these was used in 17 of the 23 studies.

The GHQ, the HADS and the STAI had problems with language, content relevance, and content coverage in studies of false positives. These instruments should not be used to measure psychological consequences of any kind of cancer screening. The PCQ was found to be the most adequate questionnaire for capturing short-term psychosocial consequences of screening mammography. However, there was little evidence that the PCQ was able to adequately detect all long-term psychosocial consequences of screening mammography.

Six of the twelve original items in the negative part of the PCQ were found to be ambiguous or poorly worded and, as a result, were reformulated. This increased the number of items in the Danish version to 18 items. All of these items were found to be relevant to negative psychosocial consequences of abnormal screening mammography. However several areas were not covered. Fifteen new items were generated resulting in a draft version of a new 33-item questionnaire measuring negative psychosocial consequences of abnormal screening mammography (part 1). In the focus group interviews it was revealed that some issues could be raised both at invitation to screening, at time of screening, in the critical period and after diagnosis. These issues were covered by part 1 of the new instrument. However, other issues could be raised only after the women had been declared “free from” cancer suspicion by knowing that their abnormal screening mammography was false-positive.

To cover these specific issues of long-term psychosocial consequences of false positives thirteen items with seven response categories each were generated during the focus group interviews (part 2). In the interviews some women stated that is was difficult to choose between some of the seven response options.

In the statistical analyses of part 1 six dimensions covering: anxiety, behavioural impact, sense of dejection, impact on sleep, breast examination and sexuality were identified. One item belonging to the dejection dimension had uniform differential item functioning. Two items not fitting the Rasch models were retained because of high face validity. A sick-leave item added useful information when measuring side effects and socioeconomic consequences of breast cancer screening. Five “poor items” identified were deleted from the final version of part 1 of the instrument.

In the statistical analyses of part 2, four dimensions covering the impact on the women’s existential values, relationship with their social network, being less or more relaxed/calm, and less or more anxious about breast cancer/less or greater belief in not-having breast cancer were identified. Rasch analyses confirmed that the response categories also had statistical problems. Therefore, a future reduction from seven to five response categories resulting in a new lay out of the questionnaire is suggested.
Conclusion and perspectives

Given the inadequacy of the measurement instruments used in previous quantitative studies on psychosocial aspects of false positives, any current conclusions about the psychosocial long-term consequences of false positives must remain tentative.

Preliminary evidence for a valid and reliable condition-specific measure for women having an abnormal and false-positive screening mammography has been established. The new instrument, the Consequences of Screening – Breast Cancer (COS-BC) consists of two parts. The first part includes 27 items measuring different attributes of the same overall latent construct – the psychosocial consequences of abnormal screening mammography. Part 1 also includes an item measuring sick-leave. Part 2 of COS-BC measures long-term psychosocial consequences of false positives. This study shows that there are substantial negative psychosocial consequences associated with having an abnormal screening mammography later confirmed to be false-positive. Consequently, the number of women receiving false positives should be kept to a minimum.

Letters and folders posted at invitation to breast cancer screening should contain information on the negative psychosocial consequences arising from abnormal and false-positive screening results.

The new understanding of psychosocial consequences of screening mammography revealed in this PhD study contributes to the balance sheet of benefits and harm of breast cancer screening and should be included in the decision-making process of whether or not to implement breast screening.

It is a goal to establish a core-questionnaire from COS-BC relevant for any kind of cancer screening. Such a core-questionnaire, COS, should consist of core-items and core-scales relevant for all kind of cancer screening programmes. Specific items, depending on the specific cancer screened for could be added to this core-questionnaire. Consequently, the COS would make it possible to potentially compare the extent of the psychosocial consequences of false-positive cancer screening results in different screening programmes.
3. **Resumé (Danish summary)**

Måling af psykosociale konsekvenser ved falsk positive screeningresultater
- mammografiscreening som et eksempel

**Formål**

Det overordnede formål med denne ph.d.-afhandling har været at identificere, tilpasse og validere et spørgeskema til måling af psykosociale konsekvenser ved falsk positive screening-mammografisvar i Danmark. Afhandlingen indeholder fire videnskabelige artikler, som dækker følgende formål:

1. at gennemgå kvantitative studier, der belyser korttids- og langtidskonsekvenser af et falsk positivt svar ved mammografiscreening, samt at identificere og vurdere anvendeligheden og relevansen af de mest anvendte spørgeskemaer til måling af disse konsekvenser.
2. at oversætte og tilpasse første del af spørgeskemaet Psychological Consequences Questionnaire (PCQ) til dansk og vurdere indholdsvaliditeten af den danske version af PCQ i forbindelse med et abnormt og et falsk positivt svar ved mammografiscreening.
3. at validere statistisk et nyt specifikt spørgeskema til måling af psykosociale konsekvenser af abnormt svar ved mammografiscreening
4. at udvikle og validere et nyt specifikt spørgeskema til måling af psykosociale konsekvenser af falsk positivt svar ved mammografiscreening

**Materiale og metode**


Det spørgeskema, som umiddelbart blev vurderet til at være mest anvendeligt til måling af psykosociale konsekvenser ved mammografiscreening, blev oversat til dansk af et panel af personer, der var dobbeltsprogede i dansk/engelsk og kvalitetsvurderet af et panel bestående af lægefolk. Indholdsvaliditeten af det oversatte spørgeskema blev vurderet i fokusgrupper med i alt 34 kvinder, som alle havde fået et abnormt svar ved mammografiscreening, et svar der senere viste sig at være falsk positivt. For at teste om dette spørgeskema var acceptabelt og ikke virkede anstødeligt, blev der også gennemført 15 telefoninterviews, dels med kvinder, der var inviteret til mammografiscreening, dels med kvinder, der aldrig havde været inviteret til eller deltaget i mammografiscreening.

Efter interviewene blev en foreløbig version af et nyt spørgeskema med høj indholdsvaliditet besvaret af kvinder, der havde fået et abnormt, et falsk positivt og et normalt svar ved mammografiscreening. Item Respons Teori (Rasch modellen) og Klassisk Test Teori blev anvendt til at analysere de indsamlede data og teste spørgeskemaets pålidelighed og gyldighed herunder begrebsvaliditet, konvergens og divergens validitet, spørgeskemaets evne til at måle forskelle mellem grupper, dimensionalitet, additivitet, objektivitet, og intern konsistens.
Resultater

Treogtyve kvantitative studier om korttids- og langtidskonsekvenser af falsk positivt svar ved mammografiscreening blev fundet og vurderet. De mest anvendte spørgeskemaer var: General Health Questionnaire (GHQ), Hospital Anxiety and Depression Scale (HADS), Psychological Consequences Questionnaire (PCQ) og State-Trait Anxiety Inventory (STAI). Ét eller flere af disse spørgeskemaer var anvendt i 17 ud af de 23 studier.

De tre spørgeskemaer GHQ, HADS og STAI havde alle problemer hvad angår sprog, relevans og dækningsgrad, når de blev anordnet i undersøgelser af konsekvenser af falsk positivt svar ved mammografiscreening. Desuden syntes ingen af disse tre spørgeskemaer at være tilstrækkelige til måling af psykosociale konsekvenser ved screening uanset typen af kræft. PCQ blev vurderet som det spørgeskema, der var bedst egnet til måling af psykosociale korttidskonsekvenser ved mammografiscreening. Imidlertid blev det også vurderet, at PCQ ikke i tilstrækkelig grad ville kunne opfange alle langtidskonsekvenser af falsk positivt svar ved mammografiscreening.

Seks af 12 originale spørgsmål i den første del af PCQ var tvetydige ved f.eks. at ét spørgsmål indeholdt to spørgsmål. Ved at omformulere disse steg antallet af spørgsmål fra 12 til 18 i den danske version af PCQ. Disse 18 spørgsmål blev af deltagerne i fokusgruppeinterviewene fundet relevante til måling af negative psykosociale konsekvenser af abnormt svar ved mammografiscreening, men dækkede ikke alle psykosociale områder. For at dække alle disse psykosociale konsekvenser, som kvinderne i fokusgrupperne gav udtryk for blev 15 nye spørgsmål udviklet. Dette resulterede i et udkast til et spørgeskema med 33 spørgsmål (del 1)

Det kom frem under fokusgruppeinterviewene at nogle spørgsmål kunne stilles såvel ved invitationen til mammografiscreening, ved selve screeningen, i den kritiske periode fra abnormt svar til kendt diagnose, og efter at diagnosen var bekræftet som falsk positiv. Derimod var der spørgsmål, der kun kunne stilles efter mistanke om brystkræft var afkræftet, og der var sikkerhed for at det abnorme screeningssvar var fald.

For at dække dette specifikke område af psykosociale langtidskonsekvenser af falsk positivt screeningssvar, blev der udviklet 13 nye spørgsmål, hver med 7 svarkategorier (del 2). Under interviewene gav nogle kvinder udtryk for, at de havde svært ved at skelne mellem nogle af de 7 svarmuligheder.

De statistiske analyser af del 1 afslørede seks dimensioner: angst, adfærdsmæssig påvirkning, modløshed, sovnproblemer, brystundersøgelse og seksualitet. Ét spørgsmål, som ikke passede ind i Rasch modellen og ikke hørte til nogen af de seks dimensioner. Disse to spørgsmål blev bevaret i spørgeskemaet, da gruppeinterviewerne viste, at de begge havde høj indholdsvaliditet. Et spørgsmål om antal sygedage, gav vigtig og anvendelig information til måling af bivirkninger af screening samt socioøkonomiske beregninger af mammografiscreening. De statistiske analyser identificerede også fem "dårlige items", der foreslås fjernet fra den endelige udgave af del 1 af spørgeskemaet.

I de statistiske analyser af del 2 blev fire dimensioner afsløret: ændring af eksistentielle værdier, forhold til socialt netværk, følelse af mere eller mindre afslappende/indre ro, større eller mindre bekymring for brystkræft/større eller mindre tillid til ikke at have brystkræft. Rasch-analyser bekræftede, at der også var statistiske problemer med nogle af svarkategoriene. Derfor er det blevet foreslået at reducere antallet af svarkategorier fra 7 til 5 med en deraf følgende ændring af layout.
Konklusion og perspektiver

Dette ph.d.-studium har vist, at de tidligere gennemførte kvantitative undersøgelser har anvendt utilstrækkelige spørgeskemaer til måling af psykosociale aspekter ved mammografiscreening. Derfor må konklusioner vedrørende psykosociale langtidskonsekvenser af falk positivt svar ved mammografiscreening forblive tentative.


Dette studium viser, at det har betydelige negative psykosociale konsekvenser at få et abnormt svar ved mammografiscreening, som senere viser sig at være et falk positivt svar. Derfor bør antallet af falk positive screeningssvar holdes på et minimum.

De breve og informationsfoldere, der vedlægges en invitation til mammografiscreening, bør indeholde informationer om de negative psykosociale konsekvenser af at få et abnormt og et falk positivt screeningssvar.

Der er under dette ph.d.-studium fremkommet ny viden om psykosociale konsekvenser ved mammografiscreening. Denne viden og indsigt kan bidrage til afvejning af fordele og ulemper ved mammografiscreening, og bør medtages i en beslutningsproces om hvorvidt mammografiscreening skal eller ikke skal implementeres generelt i Danmark.

4. Introduction

4.1. Medical screening

The term “screening” is used for the mechanical process of separating particles of different sizes. For example, huge grids are used in coal mines to screen the coal separating coal particles by size and remove trash and pieces of rock. In 1951, the United States Commission of Chronic Illness defined screening as "the presumptive identification of unrecognised disease or defect by the application of tests, examinations, or other procedures which can be applied rapidly. Screening tests sort out apparently well persons who probably have a disease from those who probably do not. A screening test is not intended to be diagnostic. Persons with positive or suspicious findings must be referred to their physicians for diagnosis and necessary treatment". This definition was also used by the World Health Organisation (WHO) in 1968 and has been adapted into Danish in the report from the Danish National Board of Health in 1990 on criteria for medical screening programmes and in the Danish Council of Ethics’ report about medical screening from 1999 (English version from 2001).

4.2. Sensitivity, specificity and positive predictive value

The aim of medical screening is to identify disease in a healthy asymptomatic population when the disease is in a symptomless phase and therefore hopefully still curable. As implied in the definition medical screening is a gross filtration and is not a perfect process. The screening test has two qualities: the ability to find persons with disease (sensitivity) and the ability to identify healthy persons (specificity). As no screening test is perfect, the sensitivity and specificity are never both 100% perfect. Therefore, persons with disease suspicion may be healthy (false positive finding) and persons with normal screening results may have disease (false negative finding), see table 1:

<table>
<thead>
<tr>
<th>Persons with disease</th>
<th>Healthy persons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive screening test</td>
<td>A (TP)</td>
</tr>
<tr>
<td>Negative screening test</td>
<td>C (FN)</td>
</tr>
</tbody>
</table>

Sensitivity: A/(A+C)
Specificity: D/(B+D)
Positive predictive value: A/(A+B)
TP: True Positive finding, FP: False Positive finding,
FN: False Negative finding, TN: True Negative finding.

“A” symbolises persons with disease that are identified in the screening programme. In contrast, “C” is the number of persons with disease that the screening test misses. “B” is the number of healthy persons that will have an abnormal screening result. Finally, “D” symbolises the healthy persons having a negative result.

The positive predictive value of a screening test expresses how many persons with a positive screening result actually have the disease. The positive predictive value of a screening test depends on the prevalence of the disease screened for. Table 2 and 3 illustrate this phenomenon:
These examples illustrate that the performance of a screening test does not only depend on the sensitivity and the specificity of the test, but also on the prevalence of the disease. The examples also illustrate that with a tenfold increase in disease prevalence in a high risk group, the positive predictive value is not ten times better. A low positive predictive value of a test is always found in a population of low prevalence despite a high specificity. In the two scenarios above the rates of false positive results are 83.7% and 32.2% respectively, with a specificity of 98%.

### Table 2: Performance of a screening test in a general population

<table>
<thead>
<tr>
<th>Persons with cancer</th>
<th>Healthy persons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive screening test</td>
<td>4</td>
</tr>
<tr>
<td>Negative screening test</td>
<td>1</td>
</tr>
</tbody>
</table>

Sensitivity: $\frac{4}{4+1} = 80\%$
Specificity: $\frac{975}{20+975} = 98\%$
Positive predictive value: $\frac{4}{24} = 16.7\%$

### Table 3: Performance of a screening test in a high risk group

<table>
<thead>
<tr>
<th>Persons with cancer</th>
<th>Healthy persons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive screening test</td>
<td>40</td>
</tr>
<tr>
<td>Negative screening test</td>
<td>10</td>
</tr>
</tbody>
</table>

Sensitivity: $\frac{40}{40+10} = 80\%$
Specificity: $\frac{931}{19+931} = 98\%$
Positive predictive value: $\frac{40}{59} = 67.8\%$

These examples illustrate that the performance of a screening test does not only depend on the sensitivity and the specificity of the test, but also on the prevalence of the disease. The examples also illustrate that with a tenfold increase in disease prevalence in a high risk group, the positive predictive value is not ten times better. A low positive predictive value of a test is always found in a population of low prevalence despite a high specificity. In the two scenarios above the rates of false positive results are 83.7% and 32.2% respectively, with a specificity of 98%.

### 4.3. Screening criteria

In 1968 WHO published a report with ten criteria that should be checked when a screening programme is to be implemented or evaluated. These criteria were mostly based on the paradigm of screening programmes for infectious diseases e.g. pulmonary tuberculosis. Therefore, one of the major concerns was the occurrence of false negative screening results, as unidentified persons with disease would continue to be a source of infection. In the last fifteen to twenty years screening programmes for cancers have been implemented increasingly in the industrialised countries. This was one of the reasons why the Danish National Board of Health (DNBH) in 1990 updated the criteria for medical screening programmes. The report clarified the WHO’s ten criteria and added four more criteria.

In 1999 the Danish Council of Ethics published a report on screening in order to emphasise that the previous suggested criteria from the WHO and the Danish National Board of Health were important and that the council wanted to contribute to the ongoing debate of medical screening.
In 1994 the EU published a report which is in line with the recommendations of the WHO and the Danish National Board of Health. The fourteen criteria of the WHO and the Danish National Board of Health are listed below:

1. The condition sought should be an important health problem (WHO)
2. There should be an accepted treatment for patients with recognised disease (WHO)
3. Facilities for diagnosis and treatment should be available (WHO)
4. There should be a recognisable latent or early symptomatic stage (WHO)
5. There should be a suitable test or examination (WHO)
6. The test should be acceptable to the population (WHO)
7. The natural history of the condition, including development from latent to declared disease, should be adequately understood (WHO)
8. There should be an agreed policy on whom to treat as patients (WHO)
9. The cost of case-finding (including diagnosis and treatment of patients diagnosed) should be economically balanced in relation to possible expenditure on medical care as a whole (WHO)
10. Case-finding should be a continuing process and not a “once and for all” project (WHO)
11. Before making any decision on the initiation of a screening activity, the following must be evaluated (DNBH):
   - validity of the testing system
   - technical efficiency
   - predictive values of test results
12. An evaluation must have been made of (DNBH):
   - the ethical and psychological consequences for the examinees
   - stigmatisation
   - the consequences of “false positive” and “false negative” test results
13. An economic evaluation must be performed (DNBH):
   - cost-benefit, cost-effectiveness, and/or cost-utility analysis
   - cash-economic evaluation
   - marginal economic evaluation
   - cost-effectiveness
14. There must be a detailed description of (DNBH):
   - programme organisation
   - steering committee (make-up, competence)
   - registration system
   - triage planning
   - provision of information to the target group
   - staff training
   - test result dissemination

This thesis contributes to an understanding of the twelfth criterion concerning ethical and psychological consequences of screening, with particular reference to the consequences of false-positive screening results.
4.4. Breast cancer screening
In industrialised countries, breast cancer is the most frequent cancer occurring in women. For a Danish woman the lifetime risk of having breast cancer is approximately 10%, and this is one of the highest incidences in the European countries. Screening mammography programmes have been suggested as a method to prevent deaths from breast cancer with the EU recommending a biannual screening test for women aged 50 – 69 years.

4.5. Sensitivity, specificity and positive predictive value of screening mammography
The sensitivity of a screening mammography for asymptomatic women ranges from 71% to 96% for one year intervals and from 56% to 86% for biannual screening. The range of the specificity is 94% to 97%. The sensitivity and specificity and thereby the positive predictive value of a screening mammography depends on factors other than the quality of the X-ray machine and the experience of the radiograph and radiologist, notably:

- the mammographic breast density, which explains why the sensitivity and specificity are lower among younger women and women taking hormone replacement therapy;
- the specificity is poorer among obese women who have a higher rate of false-positive results compared with underweight and normal weight women.

The sensitivity, specificity and positive predictive value also differ between screening clinics and countries, with the reported positive predictive value of a screening mammography ranging from 4% to 22%. On an individual level there is a large variation in the likelihood of having a false-positive screening mammography. By the ninth screening mammography the likelihood “can be as low as 5% for women with low-risk variables and as high as 100% for women with multiple high-risk factors”.

4.6. Effect of breast cancer screening
The latest meta-analysis on breast cancer screening concluded that “In the randomized, controlled trials, (screening) mammography reduced breast cancer mortality rates among women 40 to 74 years of age. Greater absolute risk reduction was seen among older women. In addition, each screening has associated risks and costs”. The number needed to screen (NNS) to prevent one death from breast cancer after approximately 14 years of observation was 1224 (NNS 665 - 2564) for the age group 40-74. Among women younger than 50 years of age, the number needed to screen in 14 years was 1792 (NNS 764 – 10540). The Swedish randomised control trials of screening mammography found a significant 29% relative risk reduction of the breast cancer mortality among women aged 50-69 years. Expressed in number needed to screen the Swedish trials showed that 1,000 women were needed to be screened biannually for ten years in order to prevent one death from breast cancer.

4.7. False-positive screening mammography
There is a large variation in the positive predictive value of a screening mammography across screening centres, national borders and from one individual to another. Therefore, the estimated lifetime risk for having a false-positive screening mammography (hereafter referred to as a false positive) also varies. The lifetime risk for a woman to have a false positive in Denmark if she participates in all ten screening rounds during 20 years from the age of 50 – 69 has been estimated to be 20-25%. In other words, every fourth to fifth woman participating in all screening rounds in
a Danish breast cancer screening programme will receive a false-positive result. The lifetime risk for having a false positive in other countries is given in table 4.

Table 4. The lifetime risk for having a false positive in other countries

<table>
<thead>
<tr>
<th>Country</th>
<th>Age group</th>
<th>Lifetime risk of a false positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>The United States</td>
<td>40 – 69</td>
<td>49.1% after ten mammographies in ten years</td>
</tr>
<tr>
<td>The United States</td>
<td>40 – 69</td>
<td>43.1% after nine screening rounds</td>
</tr>
<tr>
<td>Australia</td>
<td>50 – 69</td>
<td>37.5% biannual screened in ten rounds</td>
</tr>
<tr>
<td>Spain</td>
<td>50 – 69</td>
<td>32.4% biannual screened in ten rounds</td>
</tr>
<tr>
<td>Norway</td>
<td>50 – 69</td>
<td>20.8% biannual screened in ten rounds</td>
</tr>
</tbody>
</table>

While the rate of true positive result is the same for the United States and the United Kingdom the rate of false positives is twice as high in the USA compared to the UK.\textsuperscript{27} One of the reasons for this difference is probably the different ways the programmes of breast cancer screening are organised in the two countries.\textsuperscript{28} Another possible reason is the age group screened in the two countries: in United Kingdom women aged 50 to 65 are offered screening, while women in the United States aged 40 to 69 are screened.

4.8. Breast cancer screening in Denmark

Breast screening programmes were implemented in the city of Copenhagen in two rounds: in 1991 in the municipality of Copenhagen and in 1994 in the Frederiksberg municipality. In the county of Funen a breast cancer screening programme was implemented in 1993. However, no Danish pilot studies on psychosocial consequences of false positives were conducted prior to the implementation of these programmes. Nor were any studies conducted during programme implementation. This is probably one of many reasons for the conclusion in the report from the Danish Council of Ethics from 1999: “The Danish Council of Ethics takes a concerned view of the current situation, in which most screening programmes are initiated on an inadequate basis with too great a degree of haphazardness in the decision-making process. The National Board of Health, Denmark’s, guidelines – and WHO’s recommendations – are not being followed on a number of central points, and there is a lack of knowledge and clarity surrounding important factors relating to social and psychological effects, false negative and false positive test results, priority-setting bases and information. The debate on the possible introduction or expansion of screening programmes is often characterised by that same inadequacy.”

4.9. Other harmful consequences of breast cancer screening

Besides false positives other harms of breast cancer screening may occur such as:

- over-diagnosis and over-treatment
- identification of carcinoma in situ
- false-negative screening mammography\textsuperscript{29}
4.9.1. Over-diagnosis and over-treatment

Over-treatment is inextricably linked to over-diagnosis and over-diagnosis is an inevitable side effect of medical screening. Over-diagnosis in screening mammography may occur in three different scenarios:

1. diagnosis of breast cancer
2. diagnosis of carcinoma in situ
3. diagnosis of benign lesions that primarily are supposed to be malign

Re 1) The cancers identified in screening have an overrepresentation of slow growing cancers. Identifying a slow growing cancer in an older person that might die from other reasons a short time after screening will result in over-diagnosis. In the prevalence round a peak of cancer incidence is always seen. This peak will correspondingly occur with a decline in incidence after screening has stopped. Observational studies have reported an increase in the incidence of breast cancer and two of these also without the corresponding decline in incidence among the women older than 69 years old that were no longer screened. A follow-up study 15 years after the end of the randomised Malmö mammographic screening trial has shown an over-diagnosis of 10%.

Re 2) Approximately two third of the women diagnosed with carcinoma in situ can be seen as being over-diagnosed, because approximately only one third of carcinoma in situ lesions will progress to invasive cancer. Treatment of carcinoma in situ can therefore in some cases be regarded as over-treatment (see also the paragraph about carcinoma in situ below).

Re 3) Some of the women having an abnormal screening mammography later confirmed to be false positive will have further examination before declared free from cancer suspicion. These invasive procedures (needle biopsies and surgery) can be regarded as over-treatment. However, a woman having mastectomy as a result of a false positive has been reported.

4.9.2. Carcinoma in situ

Carcinoma in situ (CIS) is a condition in between being healthy and having a cancer. CIS in the breast range from a biologically aggressive lesion with a substantial risk of progression into invasive carcinoma, to a lesion with a very low malignant potential. A prospective 10 year follow-up study estimated that approximately one third of CIS lesions would progress to invasive cancer. In 110 consecutive, medicolegal autopsies of women aged 20 – 54 years old, 20 (18%) had CIS and 9 of those had bilateral CIS-lesions. Prior to the implementation of breast cancer screening very few in situ carcinomas were diagnosed. However, after screening had been implemented 11% of the screen-detected malignant lesions were carcinoma in situ. In the prevalence round of the Copenhagen screening programme 12% of the women diagnosed with malignant lesions had CIS and 11% in the first incidence round. In the United States, where annual screening is common and women younger than 50 are also screened, 20-25% of the malignant lesions were carcinoma in situ. Treatment of CIS is controversial and ranges from excision only, to drug treatment, to excision with radiation therapy, through to mastectomy.

4.9.3. False-negative screening mammography and interval cancer

In the European guidelines for quality assurance in mammography screening false-negative screening mammography and interval cancer has been classified as seen in table 5.
Table 5. Classification of false-negative screening mammography and interval cancer

<table>
<thead>
<tr>
<th>Categories</th>
<th>Subtypes</th>
<th>Screening films</th>
<th>Diagnostic mammogram</th>
</tr>
</thead>
<tbody>
<tr>
<td>True interval</td>
<td></td>
<td>Negative</td>
<td>Positive</td>
</tr>
<tr>
<td>Minimal signs</td>
<td></td>
<td>Minimal signs</td>
<td>Minimal signs or positive</td>
</tr>
<tr>
<td>False negative</td>
<td>Reading error</td>
<td>Positive</td>
<td>Positive</td>
</tr>
<tr>
<td></td>
<td>Technical error</td>
<td>Negative (for technical reasons)</td>
<td>Positive</td>
</tr>
<tr>
<td>Unclassifiable</td>
<td></td>
<td>Any</td>
<td>Not available</td>
</tr>
<tr>
<td>Occult</td>
<td></td>
<td>Negative</td>
<td>Negative</td>
</tr>
</tbody>
</table>

It might have negative psychosocial consequences to have a false-negative screening mammography. It depends on the time of the diagnosis. If the woman is diagnosed just after the screening it seems to have the greatest psychosocial impact, while the impact is less when diagnosed with breast cancer more than a year after screening.47 Another possible side effect of a false-negative screening result is that it might lead to treatment delay and therefore an advanced clinical stage of the breast cancer when diagnosed.26

4.10. The biopsychosocial model and cancer screening
In the 1970s the biomedical model was considered too reductionistic. The need for a new medical model resulted in the emergence of the biopsychosocial model.48 It was argued that, to provide a better basis for understanding the determinants of health and disease a medical model should take into account both the context in which the person lives and the person’s health/illness. The psychosocial aspects of the new model includes psychological, social, behavioural and cultural factors.48 The model encompasses both the psychosocial dimension and the biomedical dimension.48;49 It is not a one-way model - it is dynamic. In the biopsychosocial model people are not regarded as passive, because they are not only influenced by their surroundings but also interact with, and have an influence on, their environment. People are actively involved in their environment; they act and react if diagnosed with a disease.50;51 In a situation where a healthy person is screened for cancer and a suspicion of cancer is raised with an abnormal screening result this may have an impact on the person’s life. In Jill Cockburn’s work on psychological consequences of screening three domains were suggested to cover the psychological aspects; emotional, social and physical. A deficit in functioning in one or more of the domains was regarded as a negative consequence and hence, detrimental.52 In cervical cancer screening psychosexual and psychosocial consequences have been described after an abnormal screening test. Most notably, anxiety, stress, emotional, psychological and sexual side-effects have been reported.53-55 In this thesis the term “psychosocial” will be used according to the biopsychosocial model encompassing psychological, emotional, sexual, social, behavioural, physical and cultural aspects. For example, if anxiety reaches a level that has a significant impact on a person resulting in dysfunction it is regarded as a negative consequence and detrimental for the person’s health.

4.11. International studies on consequences of false positives
The first quantitative study on “psychiatric morbidity associated with screening for breast cancer” was published in 1989.56 Since then more than 20 quantitative studies have been published internationally covering different aspect of consequences of false positives.57 The emotional reactions and the women’s own experiences of having a false positive have also been explored in
one international qualitative study.\textsuperscript{58} This study showed that having a false positive can cause negative psychosocial consequences such as anxiety and fear. It is likely that attitudes and values towards medical screening programmes vary from one culture to another. Therefore, it seems logical to investigate screening programmes in a national context.

4.12. Danish studies on consequences of false positives

The first Danish publication on the possible psychosocial consequences women could experience when having a false positive was a qualitative study published in 1997.\textsuperscript{59} Nineteen women were interviewed, ten of whom had received a normal screening mammography and nine, a false positive. The conclusion was that women might experience negative psychosocial consequences when having a false positive. A review on psychological consequences of breast cancer screening among healthy women was published in 2000 by a Dane,\textsuperscript{60} but no Danish studies except for the qualitative study mentioned above were included. The Danish Agency of Health Technology Assessment has published a report of the breast cancer screening programme in the county of Funen covering the period 1993 - 1997.\textsuperscript{47} The report included a qualitative study where women having a false positive were interviewed. The qualitative part of this report was also concluded that false positives might have negative psychosocial consequences.

4.13. Why and how to measure the impact of false positives?

Numerically, false-positive screening results are the largest harm of screening. The number needed to screen biannually over ten years to prevent one death of breast cancer among women aged 50-69 has been found to be 1,000 women.\textsuperscript{20,21} Biannual mammography screening of 1,000 women for ten years will result in more than 100 false-positive screening results.\textsuperscript{22} The approximate proportion between prevented deaths versus false positives is therefore 1:100. The different aspects of psychosocial consequences of false positives are not measurable as in diagnostic tests (e.g. laboratory tests, radiology etc.) or by other technical means (e.g. blood pressure, weight, lung capacity etc.). Instead, other methods must be sought for their assessment. To measure comprehensively the harm associated with receiving a false positive, assessment instruments ought to capture:

- the quality of psychosocial consequences
- the extent of psychosocial consequences
- the changes in psychosocial consequences over time

4.14. The use of generic questionnaires

Generic questionnaires are designed to be used in any disease population and cover a broad aspect of the construct measured. Condition-specific instruments have been especially developed to measure those aspects of outcome that are of importance for a particular population.\textsuperscript{61} Generic measures may or may not be relevant depending on the target population and the setting in which the instrument is used. One generic questionnaire will rarely be sufficient to address the whole area covered by a research question. To overcome the lack of content coverage researchers often use a battery of questionnaires.\textsuperscript{62-68} Using more questionnaires concurrently in a survey may present a number of problems such as:
• redundancy, because items measuring the same matter or issue are repeated in different questionnaires
• overloading the respondent with numerous questions
• asking irrelevant questions
• offering the respondent different response patterns and options, which can make it difficult and confusing to respond validly\textsuperscript{61}

The advantage of widely used generic questionnaires is that results from different surveys comparing the impact of different health conditions may be comparable. This presupposes that the psychometric properties of the questionnaire are the same in the different settings. In fact, statistical re-evaluation of generic questionnaires in a new setting does not always confirm the previously estimated properties.\textsuperscript{69} Furthermore, the advantage of using generic questionnaires will be less compared to the disadvantages if the content of the questionnaire does not validly covers the research question.

4.15. Generic questionnaires in cancer screening
The psychosocial consequences of false-positive cancer screening results have been investigated in numerous questionnaire surveys reported in a recently published literature review.\textsuperscript{67} The studies had predominantly used generic questionnaires.\textsuperscript{70} Some of the frequently used measures and health status scales were: the General Health Questionnaire,\textsuperscript{71} the SF-36,\textsuperscript{72} the Hospital Anxiety and Depression Scale,\textsuperscript{73} and the State Trait Anxiety Inventory.\textsuperscript{74} A finding in the review was: “The overwhelming majority of research indicated that short-term effects were transient.” However, the validity of the measures was not evaluated. If instruments were inadequate in the setting of false-positive screening then the results and conclusions of the studies would be highly questionable.\textsuperscript{75} The question: “Assessing psychosocial/quality of life outcomes in screening: how do we do it better?” has recently been raised.\textsuperscript{76} To improve the quality of the measurement of psychosocial outcome when screening for a disease the questioner suggested that three conditions should be respected:

• the need for a control group (preferably created by randomisation)
• the need for baseline and follow up measurements
• the need for reliable measurement tools with high criterion and content validity\textsuperscript{76}

4.16. Selecting a questionnaire
It is costly and time consuming to develop and validate a new questionnaire from scratch. However, it may be possible to identify from the literature an existing validated questionnaire developed for a similar purpose.
If an apparently adequate questionnaire is identified it is likely to have been developed in another language and in a culture that may differ from the culture of the target population. In such a case it would be necessary to translate and adapt the measure. The translation and adaptation should follow an internationally accepted method including a bilingual panel and a lay panel and afterwards checked item by item in a field test with people from the target population.\textsuperscript{77} A back translation is not regarded as a sufficient quality control of a translation process.\textsuperscript{78}
Once the translation and adaptation process has been finished there is no guarantee that the psychometric properties of an adapted version will be the same as the properties of the original questionnaire.\textsuperscript{31} Statistical re-evaluations of questionnaires in new language versions have shown that previously estimated psychometric properties differ when the questionnaires are used in
different countries and cultures. It is therefore compulsory to re-validate a translated and adapted questionnaire, as if it were newly developed.

4.17. Content validity
To ensure the overall validity of a measure it is essential to ensure high content validity of a scale or a questionnaire. Content validity – including content relevance and content coverage - is the “Achille’s heel” when using generic questionnaires in contrast to using questionnaires specifically developed for a certain purpose. Therefore, after having conducted a translation and adaptation of a questionnaire it is recommended that the second step be to test the content validity of the questionnaire. The most appropriate source of knowledge about the relevance and coverage of questionnaire content is the target population. Hence, the content is best tested in either single or focus group interviews with the target population. If lack of content validity is revealed during this step it may be necessary to add new items. Deleting redundant and irrelevant items may also be necessary. If major changes are made to an existing already validated questionnaire there is only a small chance that the psychometric properties of the questionnaire will have remained constant. Therefore, a statistical re-evaluation of the modified questionnaire is needed.

4.18. Statistical methods
The statistical methods for testing questionnaire data are concentrated in two theories: Classical Test Theory and Item Response Theory.

4.18.1. Classical Test Theory
In the Classical Test Theory (CTT) normal distribution of data is assumed. It is also assumed that the responses are variables on an interval scale and that raw scores are linearly correlated. When following CTT and using for example factor analysis the purpose is to reveal one or more latent variables by exploring or confirming a linear correlation between the actual variables (the manifest variables) and the latent variable.

4.18.2. Item Response Theory
It is not necessary to make an assumption of normal distribution of data when using an Item Response Theory (IRT) model. The responses (item parameters) can be regarded as variables on an ordinal scale. In Figure 1 below, a ruler illustrates a hypothetical example of the difference between an ordinal scale and an interval scale.

Figure 1. A ruler with an ordinal scale on the top and an interval scale at the bottom.
From the ruler it can be seen that the distance between a score of 3 and a score of 4 on the ordinal scale is not the same as the distance between for example a score of 8 and 9. However, the distances between these scores (or any scores next to each other) are always the same on an interval scale. Another example illustrating the differences between an ordinal scale and an interval scale is illustrated by the item below:

<table>
<thead>
<tr>
<th>Not at all</th>
<th>A bit</th>
<th>Quite a bit</th>
<th>A lot</th>
</tr>
</thead>
<tbody>
<tr>
<td>I have felt scared.</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
</tbody>
</table>

It is not obvious that the item parameters (sometimes described as threshold parameters) in the item above are variables on an interval scale. It is actually more likely that the distances between the thresholds are unequal and the response categories are variables on an ordinal scale. In CTT the emphasis is on item parameters and their linear correlation. In contrast to the CTT, the emphasis in the IRT “is on item and person parameters, which are non-linear transformations of raw scores, and on variances of these estimates”. In the IRT the idea of unidimensionality is that a person may be represented by a single value on a single latent continuum (latent variable).

4.18.2.1. The Guttman pattern
The Guttman scale is seen as the ideal in terms of evidence of unidimensionality, figure 2.

Figure 2. The Guttman scale. Persons illustrated by letters and items by numbers.

<table>
<thead>
<tr>
<th>Items</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>D</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>E</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>F</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>G</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>H</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>I</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>J</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

In Figure 2, person A to person J represents ten persons completing a questionnaire and item 1 to item 10 represents ten items in the questionnaire. The response “1” denotes that the item is affirmed and “0” denotes that the item is not affirmed. For example, if the 10 items in Figure 2 formed an anxiety dimension then person A would be the least anxious person and person J the most anxious person. Item 1 would be the item measuring the mildest aspects of anxiety and item 10 the most severe aspects of anxiety. The unidimensionality of the Guttman scale is of absolute importance because the response pattern is perfect.

In an IRT-model the latent variable (in this case the trait of anxiety) is a hypothesis and unidimensionality is of relative matter and thereby the response pattern in an IRT-model is relative. Figure 3 illustrates such a hypothetical example of an anxiety scale.
Figure 3. A hypothetical example of ten persons completing an anxiety scale of ten items.

<table>
<thead>
<tr>
<th>Items</th>
<th>2</th>
<th>5</th>
<th>4</th>
<th>6</th>
<th>1</th>
<th>3</th>
<th>10</th>
<th>9</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>J</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>E</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>D</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>F</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>G</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

In Figure 3, person I is most anxious of all ten persons and item 2 is the item measuring the mildest aspects of anxiety. All the responses marked with yellow are responses that are not perfect compared to the Guttman pattern. The IRT-model “allows” a certain illogical and non-perfect response pattern illustrated in figure 4 as the green zone.

Figure 4. A hypothetical example of ten persons completing an anxiety scale of ten items.

<table>
<thead>
<tr>
<th>Items</th>
<th>2</th>
<th>5</th>
<th>4</th>
<th>6</th>
<th>1</th>
<th>3</th>
<th>10</th>
<th>9</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>J</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>E</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>D</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>F</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>G</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

The likelihood of responses in the green zone in Figure 4 can also be illustrated as in Figure 5 where the likelihood of affirming the items towards the severity (going right) is decreasing and the likelihood of not affirming the items increases.

Figure 5. The likelihood of responses in the Rasch model
4.18.3. The Rasch model

The Rasch model is a logistic IRT-model. The simplest Rasch model with a dichotomous outcome, has been explained using the example of jumpers jumping fences. The jumpers vary in strength from weak to strong, and the fences are of various heights posing different challenges. The probability (p) that a jumper will have a successful attempt on a fence depends on the strength of the jumper and the height of the fence. This is illustrated in Figure 6:

Figure 6. The stochastic interpretation of observations of jumpers over fences

In Figure 6, jumper A to jumper J represents ten jumpers with increasing strength and fence 1 to fence 10 represents ten fences with increasing height. The example of jumping fences can be compared with measurement of anxiety. The extent of a person’s anxiety then replaces the jumper strength in the example of Figure 6. The height of the fence is replaced by the severity of anxiety-item; a low fence is replaced by an item measuring mild aspects of anxiety and a high fence is replaced by an item measuring severe aspects of anxiety. An item measuring mild aspects of anxiety item could for example be “have you been worried about your future? yes/no” and an item measuring severe aspects of anxiety could be “have you been terrified? yes/no”.

4.18.3.1. The dichotomous Rasch model

The formula of the Rasch model with dichotomous outcome is:

\[ p(\beta) = \frac{e^{(\beta - \delta)}}{1 + e^{(\beta - \delta)}} \]

Like in the Guttman scale the Rasch model assumes that: A) the “milder” the item is, the more likely it will be affirmed; and B) the more affected the respondent is, the more likely the respondent will affirm an item compared to a less affected individual. However, the Guttman scale assumes a perfect response pattern while the Rasch model assumes a probabilistic Guttman response pattern. In the formula of the dichotomous Rasch model \( p(\beta) \) is the probability that a person who is affected to a certain degree “\( \beta \)” will affirm an item measuring the severity “\( \delta \)”. Thereby, the probability of a person affirming an item depends on how much the person is affected and the severity of the item.
If a person responds to an item where the severity of the item is exactly the same as the degree to which the person is affected then $\beta = \delta$ and $p(\beta) = 0.5$ (50%). If the severity of the item is less than the person’s level of affection ($\beta > \delta$) the probability will be greater than 50% and when $\beta < \delta$ the probability will be less than 50%.

To explain the formula of the dichotomous Rasch model and the calculation in the previous paragraph a set of dichotomous items forming an anxiety dimension is used. The anxiety-item parameters would be manifest variables on a latent anxiety-continuum most likely on an ordinal scale. A person with a certain degree of anxiety would have a certain location on such a latent continuum (latent variable). The person would affirm items that are located below the person’s location with a probability of more 50% and the probability of affirmation would increase the milder the item was (beginning in the centre of Figure 5 and going left). The person would affirm the items located above the person’s location with a probability of less than 50% and the probability of affirmation would decrease with the more severe aspects of anxiety the item is measuring (beginning in the centre and going right in Figure 5). The Rasch model takes into account that measurement error always exists, that no respondents are 100% logical in their responses and that the pattern is not perfect (Figure 2) but allows for probabilistic variations (Figure 3-5).

4.18.3.2. The advantages of the Rasch model

The Rasch model tests whether the manifest variables are fitting a probabilistic Guttman pattern and if that is the case, then the hypothesised latent variable is confirmed. If there is a misfit of the manifest variables to the Rasch model then the items having misfit to the model can be identified. After deleting the item with the most misfit the remaining items can be tested in the model. The items that do not fit the model should be treated either as single items or “poor” items depending on there face and content validity.

Where items are shown to fit a Rasch model the measure can be shown to possesses criterion-related construct validity, unidimensional, additive, specifically objective, sufficient and reliable. The Rasch model is the only IRT-model that possesses specific objectivity and sufficiency. If measurement has specific objectivity in relationship to the latent variable, then a comparison between two persons using a selected set of items that represents the range of the scale (dimension) does not differ systematically from comparisons using another set of valid items from the scale. Similarly, comparison of two persons should not depend on measurements of other persons. The sufficient statistic contains all available information about the parameter of interest. If for example, it is known how many dichotomous items that have been affirmed on a Rasch fitting scale (the score on the dimension is known) more knowledge is not gained by knowing which of the items on the scale were affirmed. The score on the scale contains all the available information.

The Rasch model is considered a valuable “gold standard” because the Rasch model:

- has no presumption of normal distribution of data
- can include data on an ordinal scale
- provides formal representation of perfect measurement, because a measure fitting the Rasch model possesses:
  - criterion-related construct validity
  - unidimensionality
  - additivity
  - specific objectivity
  - sufficiency
  - reliability
4.19. The purpose of the thesis

The overall purpose of this PhD study was to identify, adapt and validate an instrument to measure the psychosocial consequences of false-positive breast cancer screening results in Denmark. The aims of the four papers included in this thesis are:

1. to review quantitative studies on women’s experiences of short-term and long-term consequences of false positives and to identify and assess the adequacy of the most frequently used instruments for measuring these consequences (I)
2. to translate and adapt into Danish the negative part of the Psychological Consequences Questionnaire and to assess the content validity of the Danish version of the instrument in the setting of abnormal screening mammography later confirmed to be false-positive (II, III)
3. to statistically validate a new condition-specific instrument measuring psychosocial consequences of abnormal screening mammography (IV)
4. to develop and validate a new instrument measuring long-term psychosocial consequences of false positives (II, III, V)

4.20. The steps of the PhD study

The Roman numerals in the flow chart correspond to the steps of the study described in the papers.
5. **Article 1**

The adequacy of measurement of short and long-term consequences of false-positive screening mammography

John Brodersen, Hanne Thorsen, Jill Cockburn

Objectives: The aim of this study is to review quantitative studies on women’s experiences of consequences of false-positive screening mammography to assess the adequacy of the most frequently used instruments for measuring short-term and long-term psychological consequences.

Methods: Relevant papers reporting quantitative studies on consequences of false-positive screening mammography were identified using MEDLINE, CINAHL, EMBASE and PsycInfo databases. Articles citing development and psychometric properties of the most frequently used measures were also retrieved. Finally, the review focused on studies that had used at least one of the most frequently used measures.

Results: Twenty-three relevant studies were identified. The most commonly used measures were the General Health Questionnaire (GHQ), the Hospital Anxiety and Depression Scale (HADS), the Psychological Consequences Questionnaire (PCQ) and the State-Trait Anxiety Inventory (STAI). One or more of these was used in 17 of the 23 studies.

Conclusions: The GHQ, the HADS and the STAI have problems with language, content relevance, and content coverage in studies of false-positive screening mammography. These instruments should not be used to measure psychological consequences of any kind of cancer screening. The PCQ is an adequate questionnaire for measuring short-term consequences, and the PCQ is preferable to other measures because of its higher sensitivity. However, there is little evidence that the PCQ is able to adequately detect all long-term consequences of screening mammography. Given the inadequacy of the measurement instruments used, any current conclusions about the long-term consequences of false-positive results of screening mammography must remain tentative.

INTRODUCTION

Specificity of biennial screening mammography has been reported as 97%, with positive predictive values of 9% and 17% for women aged 50–59 and 60–69, respectively. Consequently, 10 out of 11 women in their fifties and five out of six in their sixties receiving an initial abnormal screening mammography do not have cancer. Thus, an abnormal screening mammography more often provides a false-positive than a true-positive result. In the UK, where three-yearly screening is offered to women aged 50–65, more than 50,000 women per year will receive a false-positive mammography.

The WHO recognise that cancer screening may have a negative affect on the target population, even when their ‘Ten Commandments’ for medical screening are adhered to. Gray and Austoker are more direct about the potential negative affects of screening: ‘All screening programmes do harm; some also do good.’ The ultimate damage caused by a screening program arises when healthy people die, either from diagnostic procedures or from medical interventions following a false-positive result. This has been reported after ovarian cancer screening, colorectal cancer screening, and lung cancer screening. In breast cancer screening, a case of suicide following receipt of a recall letter has been reported.

Mortality data indicate some benefits of cancer screening programs. It is possible to calculate financial burden and time consumption of screening on participants, medical practice and society. However, adverse effects such as psychosocial impact of a false screening result on the individual, are difficult to define and measure.

The aim of this paper is to review quantitative studies on women’s experiences of consequences of false-positive screening mammography to assess the adequacy of the most frequently used instruments for measuring short-term and long-term consequences.

METHODS

Relevant papers reporting quantitative studies on consequences of false-positive screening mammography were identified using MEDLINE, CINAHL, EMBASE and PsycInfo databases. Reference lists were manually searched for articles not identified through the electronic means. This process was continued until no more relevant papers were identified. Articles citing development and psychometric properties of the most frequently used measures were also retrieved. Finally, the review focused on studies that had used at least one of the most frequently used measures. The adequacy of measures was based on whether field-tests and psychometric analyses were appropriately conducted in the setting of false-positive screening mammography.

RESULTS

The search identified 30 papers (including three reviews), covering 23 studies measuring the consequences of false-positive mammography. The search also identified five papers dealing with women’s attitudes, experiences, satisfaction and consequences of attending a breast screening program, one pilot study on the same issues, and one study of false-positive thermography.
Fifteen studies investigated only short-term consequences of false-positive mammography i.e. up to three months after women were informed that they did not have breast cancer.12–17,19–22,27,32,34–38 All reported adverse short-term consequences.

Seven studies measured short-term and long-term consequences (over more than three months).18,23–26,28,29,31,33 They also consistently report adverse short-term consequences but results regarding long-term consequences are ambiguous. One study found that 15% of the women receiving a false-positive mammography had measurable adverse consequences after six months.28 Brett et al. found adverse long-term consequences at five, eleven, and 35 months follow up.25,31 and three studies found no long-term consequences.18,26,33 Interpretation of results in two of the seven studies is unclear.24,29 One study measured only long-term consequences and found that they were present after two years.30

In the 23 different studies at least fifteen different measures have been used. Those used most often were the General Health Questionnaire (GHQ), the Hospital Anxiety and Depression Scale (HADS), the Psychological Consequences Questionnaire (PCQ) and the State-Trait Anxiety Inventory (STAI). One or more of these was used in 17 of the 23 studies. Table 1 summarises the results of the review. Item generation and scale development, scoring system, psychometric properties and results of measuring consequences of false-positive screening mammography are reported for each of the four scales.

### General Health Questionnaire (GHQ)

The GHQ is a self-administered questionnaire designed as a screening tool of psychiatric disorders in non-psychiatric clinical settings such as primary care or general medical out-patients.31 The original version, developed in 1972, contained 60 items covering four areas; depression, anxiety, objectively observable behaviour and hypochondriasis. The four dimensions arose mainly as a result of interviews conducted with 2460 non-hospitalised adults about aspects of adjustment and distress in 196032 and from interviews using the Cornell Medical Index31,34 with 120 residents of a Jewish housing project in 1965. Discussions were also conducted with experienced psychiatrists.31 The anxiety scale used Fried and Lindemann’s work of ‘role-satisfaction’.35 Items were also taken from other scales and inventories and 30 items were developed specifically for the GHQ.31

Items enquire about recent experience of symptoms or behaviour, with four response options (‘less than usual’, ‘no more than usual’, ‘rather more than usual’, and ‘much more than usual’) scored 0-1-2-3, as an adjectival scale on a con-

### Table 1 Quantitative studies on consequences of false positive mammography, reviews, and related papers. The articles are listed by year of publication and include two qualitative studies on consequences of false positive mammography.29,52

<table>
<thead>
<tr>
<th>Authors</th>
<th>Publication year</th>
<th>PCQ</th>
<th>GHQ</th>
<th>HADS</th>
<th>STAI</th>
<th>Other questionnaires/items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ellman et al.</td>
<td>1989</td>
<td></td>
<td>GHQ-28</td>
<td></td>
<td></td>
<td>Own questionnaire</td>
</tr>
<tr>
<td>Gram et al.</td>
<td>1990</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bull et al.</td>
<td>1991</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Self conducted telephone interview</td>
</tr>
<tr>
<td>Lerman et al.</td>
<td>1991</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Self conducted telephone interview</td>
</tr>
<tr>
<td>Lerman et al.</td>
<td>1991</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Own questionnaire</td>
</tr>
<tr>
<td>Gram et al.</td>
<td>1992</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cockburn et al.</td>
<td>1994</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Coping test and MACL</td>
</tr>
<tr>
<td>Udbrink et al.</td>
<td>1995</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sutton et al.</td>
<td>1995</td>
<td></td>
<td>GHQ-28*</td>
<td></td>
<td></td>
<td>STAI-Y EPI &amp; CSI</td>
</tr>
<tr>
<td>Chen et al.</td>
<td>1996</td>
<td></td>
<td>GHQ-12</td>
<td></td>
<td></td>
<td>Plus one other item</td>
</tr>
<tr>
<td>Swanson et al.</td>
<td>1996</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Plus others items</td>
</tr>
<tr>
<td>Ong, Austoker &amp; Brett</td>
<td>1997</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>STAI, 6-items form</td>
</tr>
<tr>
<td>Scal-Klomp et al.</td>
<td>1997</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>SOM, Fear of Cancer</td>
</tr>
<tr>
<td>Brett, Austoker &amp; Ong</td>
<td>1998</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Plus others items</td>
</tr>
<tr>
<td>Gilbert et al.</td>
<td>1998</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>STAI-Y, anxiety-trait scale</td>
</tr>
<tr>
<td>Lowe et al.</td>
<td>1999</td>
<td></td>
<td>GHQ-28§</td>
<td></td>
<td></td>
<td>Plus others items</td>
</tr>
<tr>
<td>Olsson et al.</td>
<td>1999</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aro et al.</td>
<td>2000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>STAI-X Illness Attitude Scale, Beck’s depression scale plus own questionnaire and other items</td>
</tr>
<tr>
<td>Lipkus et al.</td>
<td>2000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>CED-S plus other items</td>
</tr>
<tr>
<td>Brett &amp; Austoker</td>
<td>2001</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Plus other items</td>
</tr>
<tr>
<td>Ekkeberg et al.</td>
<td>2001</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Plus other items</td>
</tr>
<tr>
<td>Lamping et al.</td>
<td>2001</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Numerical anxiety scale</td>
</tr>
<tr>
<td>Lindfors et al.</td>
<td>2001</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Own questionnaire plus other items</td>
</tr>
<tr>
<td>Meytre-Agustoni et al.</td>
<td>2001</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Plus one other item</td>
</tr>
<tr>
<td>Huslop et al.</td>
<td>2002</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Items on women’s satisfaction</td>
</tr>
<tr>
<td>Sandin et al.</td>
<td>2002</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>SCL-90 plus other questions</td>
</tr>
<tr>
<td>Bowland et al.</td>
<td>2003</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>SSQ, IOAT, a coping scale, plus other items</td>
</tr>
</tbody>
</table>

**Reviews:**

Rimmer & Bluman 1997
Steggles et al. 1998
von Bulow 2000

**Other Related studies:**

Dean et al. 1986 +
Bartolucci et al. 1989 +
Elkind et al. 1990 +
Olsson et al. 1993 +
Lightfoot et al. 1994 +
Walker et al. 1994 +
Ong and Austoker 1997 +

**Qualitative studies:**

Lunde 1997
Pagdett et al. 2001

* 7 items from the GHQ-28 anxiety sub-scale
§ The depression sub-scales excluded
The HADS is a patient-completed measure developed in 1983 to provide a reliable screening test for psychiatric disorders in patients in non-psychiatric hospital departments.57 The measure can also be scored dichotomously, with ‘0’ given for ‘less than usual’ and ‘no more than usual’, and ‘1’ for ‘rather more than usual’ and ‘much more than usual’.57 For the latter method, scores are added and if a person scores more than a given threshold they are classified as a ‘case’.57 A 140-item version of the GHQ was published with the original 60 items.38 Several different shorter versions have also been developed, namely GHQ-12, GHQ-28 and GHQ-30.58 The ‘case’ thresholds vary with different versions.58

The versions of the GHQ have been found to be reliable and valid, when subjected to traditional psychometric analysis such as test–retest reliability, internal, and concurrent validity in a variety of settings.56 Two studies using item response theory (latent trait model) have tested the GHQ-1258 and the GHQ-3059, respectively. Both studies found a difference between positively and negatively orientated items, indicating that the GHQ may not be a unidimensional measure and perhaps it would be appropriate to split it up into a positive and a negative component.

Five of the reviewed studies used the GHQ-12 or GHQ-28 (Table 1).12,20,21,24,27 In one study only the GHQ-28 anxiety sub-scales was used60 and in another, the depression sub-scale was excluded from the GHQ-28.27 An overall GHQ score has been calculated in three of the four studies.12,20,27 while the other classified women as ‘cases’ or ‘non-cases’.21 Lowe et al. reported ‘acceptable’ Cronbach’s alphas for the three subscales that were used from the GHQ-28.57 None of these studies have reported pre-testing or test-retest reliability in these settings.

**Hospital Anxiety and Depression Scale (HADS)**

The HADS is a patient-completed measure developed in 1983 to provide a reliable screening test for psychiatric disorders in patients in non-psychiatric hospital departments.61 The HADS has fourteen items equally distributed in two subscales: depression and anxiety. The depression subscale was developed from the anhedonic state,62 as the authors argue that this responds well to antidepressant drug treatment.62 The anxiety subscale was developed using modified items from the Hamilton Anxiety Scale (a rating scale for patients already diagnosed with neurotic anxiety)63,64 and the Present State Examination (a structured interview that assesses the present mental state of adult patients suffering from either neuroses or functional psychoses).63 All items in the PSE were selected by psychiatrists and physicians.

Each item has four response options. Subjects select the option that best describes how they have been feeling over the previous week. These are scored 0–1–2–3, as an adjectival scale on a continuum with discrete responses. Items added to give a total score.61

The HADS was first validated by comparing scores of 98 adult outpatients from a general medical clinic with ratings of clinicians blinded to patients’ scores. These were used to determine score thresholds for ‘non-cases’ (score of seven or less), ‘doubtful cases’, (eight to 10) and ‘cases’ (11 or over). Second, a selection of patients were matched for age and gender with a ‘normal’ sample to test if the HADS scores were influenced by physical illness apart from mood disorders. The HADS has been used in a range of disease areas, with approximately 500 studies identified. Few have used a healthy person comparison group. Traditional psychometric assessments, mainly in studies that involve patients, have shown acceptable results.66

The HADS has been used in five of the reviewed studies (Table 1).14,24,26,12,15 These do not report pre-testing, test–retest reliability, or analyses of internal consistency in these settings.

**Psychological Consequences Questionnaire (PCQ)**

The PCQ was developed in Australia in 1992 as a self-administered measure of psychological consequences of screening mammography.67 Items encompassing emotional, social and physical domains were generated by reviewing published scales and research. Scales reviewed included the GHQ,68 the Sickness Impact Profile,68 a scale measuring whether screening programs cause morbid concern in the community,69 and a number of oncology quality of life measures.70,71 Interviews were also conducted with women attending screening and recall for further investigations after initial screening.

A sorting procedure showed high level of agreement between expert judges as to whether the postulated physical, social and emotional dimensions could be distinguished on the basis of item content. After pilot testing with other samples of women from the screening program, five items measuring emotional issues, four measuring physical and three measuring social issues were retained.

The questionnaire asks ‘Over the last week how often have you experienced the following because of thoughts and feelings about breast cancer?’ followed by the items. Response options are ‘not at all’, ‘rarely’, ‘some of the time’, and ‘quite a lot of the time’, scored from 0–3. Ratings for items within each dimension are added to give respective subscale scores. A higher score indicates greater dysfunction on that dimension. A pilot version of the PCQ was field-tested in screening and recall clinics.57 Items were tested for floor effect, ability to measure differences between groups of women and women’s perceptions of ease of completion.

Responses of women at initial screening and assessment clinics were compared with ratings of a clinical interviewer, who was blinded to the PCQ score. The high level of agreement for each subscale was found as an indicator for concurrent validity. Women’s initial scores (obtained at the screening clinic) were compared with scores obtained during the recall visit. As scores varied in the predicted direction (higher scores, indicating greater dysfunction were found at the recall clinic) this was taken as evidence for construct validity.

The PCQ has been used in ten of the reviewed papers (Table 1).18,22,23,25,27,28,31,15,16,18 A Swedish version of the PCQ together with the Symptom Check List (SCL-90) was completed by 220 women in a pilot study but data on concurrent validity were not reported.28 PCQ has been shown to be more sensitive than a 6-item short form of the STAI,23 the anxiety-trait scale of the STAI-Y,23 and the GHQ-28 (excluding the depression subscale).27 Three studies have performed factor analyses on the PCQ. Ong et al. found a single factor with an eigenvalue of 7.7, explaining 65% of the variance.72 Olsson et al. did not report their results, other than saying that the ‘analysis did not suggest three underlying factors’.72 Hislop et al. also do not report results but state that two dimensions were found.15 Cronbach’s alpha values are presented in Table 2.

**State-Trait Anxiety Inventory (STAI)**

The development of the STAI was initiated in 1964 with the purpose of producing ‘an objective, self-reported research instrument that could be used to measure both state and trait
anxiety in normal adults. A large number of items from three widely used anxiety-trait scales were reformulated so that each could be used as a measure of both anxiety-state and anxiety-trait. A number of tests were performed with college and university students, who were asked either to imagine hypothetical situations such as being relaxed or at being at an examination or by giving the questionnaire to the students in a period of non-examination and later in a period of examination. The result was a single scale, the Form-A of the STAI (STAI-A).

Psycholinguistic problems led to the development of the STAI-X, which comprises a 20-item anxiety-trait scale (describing how the respondent feels in general) and a 20-item anxiety-state scale (describing how the respondent feels at a particular moment). Further refinement after testing with high school students led to the 40 item STAI-Y, where weaker items from STAI-X were replaced either by equal or better psychometric items or by items that were more consistent with the concept of state and trait anxiety.

There are four response options for each item in all the forms of the STAI, scored on a scale of 1-2-3-4 as an adjectival scale on a continuum with discrete responses. The higher score the higher the likelihood of either state or trait anxiety. Reliability and validity of the many forms of the STAI have been tested in several studies using traditional psychometric assessments. Results are acceptable. One study reports analyses based on item response theory on STAI-X data. Rasch analyses indicated that several items in both scales of the STAI-X ‘produce misfit responses, share identical locations on the continuum, and do not produce equal units of measure’. Furthermore, the STAI-X was not accurate enough to differentiate the trait and state anxiety levels. Only ten items of the trait-scale and nine from the state-scale were applicable, and there was a lack of items covering the measures of ‘light’ and ‘severe’ trait and state anxiety.

The STAI has been used in five of the reviewed papers but each study used a different version (Table 1). Sutton et al. used the STAI-Y; Ong, Brett and Austoker used a 6-item short form of the STAI-Y in their pilot study and in a later study they used the anxiety-trait scale of the STAI-Y. Aro et al. used the anxiety-state scale from the STAI-X and Hislop et al. do not report which version of the STAI was used. The five studies in this setting did not report pre-testing or test-retest reliability. Only one of the five studies reports analyses of internal consistency (Cronbach’s alpha 0.96).

**DISCUSSION**

The review has uncovered a number of issues with existing measures that cast doubts about their overall adequacy for measuring consequences of false-positive screening mammography. Firstly, the language used in items may not be relevant for current use. The GHQ, the HADS, and the STAI originate from the 1970s and 1980s and their item generation goes as far back as the 1950s. The language of a questionnaire must be kept up to date as the linguistic value of words and terms can take on new meanings over time. Both the wording of the items and the construct behind the measures could be different if the measures had been developed more recently.

Secondly, the content of items in some measures is not relevant for women attending for breast cancer screening and who may experience a false-positive mammography. The GHQ was developed to tap psychiatric disorders in non-psychiatric clinical settings and the HADS to tap psychiatric disorders in non-psychiatric hospital departments. Middle-aged healthy women participating in breast cancer screening and experiencing false-positive screening mammography can hardly be categorised as suffering from psychiatric disorders as such. Therefore, it is questionable whether the GHQ and the HADS can accurately capture the consequences of such an experience.

The STAI was developed to tap state and trait anxiety in normal adults. However, it is doubtful that the anxiety experienced by students before an examination equals the anxiety experienced by women threatened by breast cancer. Furthermore the lack of items in the STAI covering ‘light’ and ‘severe’ state and trait anxiety, shown by Rasch analyses, throws doubt on the STAI’s ability to capture the full construct of anxiety. Moreover as the GHQ, the HADS, and the STAI cover psychiatric and psychological disorders in a broader sense, they may impose irrelevant items on women experiencing false-positive mammography. Streiner and Norman have warned that if items appear irrelevant, the respondent may object to them and omit them, irrespective of the instruments ‘possibly superb psychometric properties’. Using more than one such measure could exacerbate the perception of irrelevance.

Thirdly, the variety of additional measures used in the 23 studies included in this review (Table 1) shows that the researchers acknowledge the necessity of supplementing the overall psychological and psychiatric measures to ensure content coverage. However, such variation makes it difficult to compare studies. Response burden is another disadvantage. It would be preferable if the same level of information could be achieved by using a short questionnaire specifically developed to cover a target.

Despite the different questionnaires used, all studies found adverse short-term consequences of false-positive screening mammography. This indicates at least some relevance of the batteries of measures chosen by the researchers. However, it does not indicate content coverage. Surprisingly, there is no evidence to suggest pre-testing of content relevance and coverage of the GHQ, the HADS and the HADS.
the STAI in any of the studies using these measures in this setting. If pre-testing had been conducted it might have shown that these measures do not cover all short-term and long-term consequences of false-positive mammography.

Fourthly, in all the reviewed studies little attention has been paid to the applicability of the measures. Cockburn et al. first contributed to this discussion and developed a new questionnaire (PCQ) specific to the area. Sutton also addresses the applicability of measurement, and acknowledges that there could have been a different pattern of results in his study if a more specific instrument such as the PCQ had been used. The better sensitivity of the PCQ to detect psychological consequences of false-positive mammography compared with a six item short form of the STAI, the anxiety-trait scale of the STAI-Y and the GHQ-28 exclusive the depression subscale, also suggests that the PCQ is more applicable to this setting than other instruments.

Fifthly, there are few assessments of the psychometric properties of instruments in the setting of false-positive consequences of screening mammography. Only two studies have reported analyses of internal consistency of three of four subscales of the GHQ-28 and an unknown version of the STAI in this setting. However, the relevance of adequate internal consistency is questionable when there has been no examination of the measure’s content validity in the setting. In contrast, the consistently high levels of internal consistency of the PCQ that have been reported in various studies, along with the demonstrated content validity of the measure (at least for short-term consequences) gives further confidence of the adequacy of the PCQ for this setting.

Lastly, there is insufficient evidence that any of the current measures adequately describe the longer term consequences of receiving a false-positive result. This could explain the ambiguity of results in this area. The PCQ was used in two of the three studies that found some long-term consequences. It appears that the method of development of the PCQ ensured that at least some items are relevant for consideration of long-term consequences. However, development of PCQ items did not involve women for whom some time had passed since receiving their false-positive result. Thus, there may be other relevant long-term consequences that have been missed. For example, Gram et al. report that one long-term consequence described by women after diagnostic surgery, was reduced sexual sensitivity of the breast.

Women also mention sexual consequences of false-positive mammography in a Canadian study about women’s attitude to breast screening. Measuring consequences of screening mammography 43

CONCLUSIONS

This review has revealed that the GHQ, the HADS and the STAI have problems with language, content relevance, and content coverage when these measures are used in studies of false-positive screening mammography. Given that the adequacy of these measures in this setting is doubtful, it is suggested that these instruments should not be used to measure psychological outcomes related to the consequences of any kind of cancer screening. It is worrying, therefore, that studies in colorectal screening are using these measures.

The review suggested that the PCQ is an adequate questionnaire for measuring short-term consequences of receiving a false-positive mammography and that the PCQ is preferable to other measures because of its higher sensitivity. However, the review also revealed that this instrument has some deficiencies when measuring long-term consequences of screening. There is little evidence that the PCQ is able to adequately detect all long-term consequences of screening mammography. The items were developed on women who were waiting to hear about their false-positive result, and did not include women who were told that their abnormal screening mammography was false. Women on early recall, women who had diagnostic surgery before being given a benign result and women whose false-positive result had occurred at some time in the past, were also not included.

This is an important issue that needs to be examined further in rigorously conducted qualitative studies that explore the range of issues for women viewing their false-positive results in the longer term, as a first step to develop a relevant measure in this setting. Until this is achieved, any conclusions about the long-term consequences of false-positive results of screening mammography must remain tentative.

ACKNOWLEDGEMENTS

We would like to thank Professor Thomas Kohlmann, University of Greifswald for his great help in interpretation of the papers about Latent Trait Theory, and Dr John Sahl Andersen, Copenhagen University, who has been an inspiration for our work with this review.

REFERENCES


6. Discussion of article 1

6.1. Major findings
The literature search identified 30 papers (including three reviews) covering 23 studies measuring consequences of false-positive screening mammography (hereafter referred to as false positives). Fifteen of the 23 studies investigated only short-term consequences of false-positive mammography i.e. up to three months after women were informed that they did not have breast cancer. All reported adverse short-term consequences. Seven studies measured short-term and long-term consequences (more than three months). While these studies all reported adverse short-term consequences, results regarding long-term consequences were ambiguous. One study measured only long-term consequences and found that they were present after two years.

In the 23 different studies at least fifteen different measures have been used. The most commonly used measures were the General Health Questionnaire (GHQ), the Hospital Anxiety and Depression Scale (HADS), the Psychological Consequences Questionnaire (PCQ) and the State-Trait Anxiety Inventory (STAI). One or more of these was used in 17 of the 23 studies. The GHQ, the HADS and the STAI had problems with language, content relevance and content coverage in studies of false positives. Therefore, these instruments are not adequate for measuring psychological consequences in cancer screening. The PCQ was found to be an adequate questionnaire for measuring psychosocial consequences of abnormal screening mammography and preferable to other measures because of the PCQ’s higher sensitivity. However, there is little evidence that the PCQ is able to adequately detect all long-term psychosocial consequences of screening mammography. The PCQ was developed in Australia and the original version is reproduced in appendix I (page 120).

6.2. Assessment of methods

6.2.1. Feasibility of the comprehensive literature search
The search strategy involved no language restriction and was conducted in four different databases: MEDLINE, CINAHL, EMBASE and PsycInfo. The following words were used in different combinations: “stress”, “nervous system physiology”, “diagnostic errors”, “behavioral disciplines and activities”, “psychological phenomena and processes”, “behavior and behavior mechanisms”, “adverse effects”, “arousal”, “adaptation, psychological”, “coping”, “adverse”, “consequences”, “false positive reactions”, “false positive”, “diagnostic errors”, “emotional”, “mammography”, “breast”, “screening”, “mass screening” “outcome assessment (health care)” and “questionnaires”. The overall search strategy for the MEDLINE database is reported in appendix II (page 122). The abstracts of the papers that emerged from the literature search were read to identify relevant papers. A relevant paper was defined as one reporting a study of psychosocial consequences in relation to breast cancer screening (or related issues; quality of life, stress, etc. according to the words used in the literature search). When a relevant paper was identified then the reference list was manually searched for articles not identified through the electronic means. This process was continued until no more relevant papers were identified.

Two non-English papers written in Danish were identified; a qualitative study and a review. Two other English reviews were identified. One of these was published prior to the Danish qualitative study. All of the papers identified by the current review were also identified by these
other published reviews with the exception of the Danish qualitative study.\textsuperscript{59} This qualitative study was also not included in a review published after the current review.\textsuperscript{57} Although many different instruments had been used the adequacy of only four measures was explored:

- The first criterion for choosing these four measures was that the development and validation of the instruments identified should have been published. Only with such publications available it would be possible to discuss the instruments adequacy in the setting of false positives.
- The second criterion was that only questionnaires used in at least two different studies were explored. If the same instrument had been used by at least two different researchers it might have been more relevant and adequate.

6.3. Update of literature searched

An updated literature search covering the period December 2003 to 2005 using the same four databases reported above has been conducted following the same search strategy (appendix II, page 122). Four questionnaire surveys on psychosocial aspects of breast cancer screening were identified and are listed in table 6:

Table 6. Quantitative studies on consequences of mass screening when screening for breast cancer published in the period December 2003 – 2005. The articles are listed by year of publication and the measures used in the studies are provided (not including sociodemographic variables).

<table>
<thead>
<tr>
<th>Authors</th>
<th>Publication year</th>
<th>Questionnaires/items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barton et al\textsuperscript{95}</td>
<td>2004</td>
<td>The Impact of Events Scale and the Hopkins Symptom Checklist anxiety and depression subscales</td>
</tr>
<tr>
<td>Domar et al\textsuperscript{96}</td>
<td>2004</td>
<td>STAI, McGill Pain Questionnaire and two VAS-scales from 1 to 10 as self-reported levels of anxiety and pain</td>
</tr>
<tr>
<td>Heckmann et al\textsuperscript{97}</td>
<td>2004</td>
<td>STAI and Ways of Coping Questionnaire</td>
</tr>
<tr>
<td>Brunton et al\textsuperscript{98}</td>
<td>2005</td>
<td>Self developed questionnaire and an anxiety level measured by a 4-point Likert scale</td>
</tr>
</tbody>
</table>

None of the studies reported tests on content validity of the instruments used or statistical explorations of the psychometric properties of the measures. The four studies have used questionnaires and items not evaluated in the present review. However, no new measures had to be explored according to the two previously mentioned criteria. The number of different questionnaires used in surveys on psychosocial aspects of screening mammography is more than 20 after this updated literature search. The four studies published in Table 6 would not have resulted in any important changes in the discussion and conclusion of the present review if it was to be updated.

6.4. Justification of conclusion

The PCQ was developed by Jill Cockburn to assess the psychological consequences of the actual act of participation in breast cancer screening. As a part of the development process, interviews were conducted with women who had received a normal screening mammography and with women who had received an abnormal screening mammography. The interviews revealed that the women experienced impact in emotional, physical and social areas.\textsuperscript{52} Impact on sleep Cockburn categorised
as a physical impact. Other studies have reported more frequent breast self-examination as a consequence of false positives. Therefore, the three generic questionnaires; the GHQ, the HADS and the STAI lack content validity in the setting of psychosocial consequences of false-positive cancer screening. None of these instruments on their own include all reported domains; emotional, physical, social, sleep, sexual and breast self-examination. The PCQ is the most preferable instrument because it covers, albeit not comprehensively, the above domains and it has been developed from interviews with some of the women from the target population. The PCQ has also been developed and validated in the setting of screening mammography.52

No qualitative study on the long-term consequences of false positives has been identified. Neither have any tests been reported on the content validity of these instruments in the setting of long-term psychosocial consequences. Therefore, there is little evidence that the PCQ (or any other measure) is able to adequately detect all long-term psychosocial consequences of false positives. Given the inadequacy of instruments in measuring the long-term psychosocial consequences no valid conclusions can be drawn on this issue from the questionnaire surveys.

6.5. Contribution to the current knowledge
The lack of content validity of generic instruments in the setting of false positives was made clear in the present review. The psychometric properties of generic measures are not necessarily the same when measures are applied in two different populations.69 Furthermore, none of the papers included in the present review reported re-evaluation of the generic instruments in the setting of screening mammography. If no generic questionnaires are found adequate a condition-specific measure should be developed and validated. After the present review was published “the need for reliable measurement tools with high criterion and content validity” was raised in an editorial on how to assess better the psychosocial consequences of screening.76 The present review was used as an argument for the use of condition-specific measures because they are more sensitive. Among criteria raised by the editorial on how to deal with false positives the issue on content validity and sensitivity of the applied instruments was especially emphasised. As described in the review the PCQ is more sensitive than the STAI and the GHQ in the setting of false positives. This is probably caused by a higher content validity of this instrument compared to the generic questionnaires.
7. **Article 2**
Psychosocial consequences of abnormal screening mammography – development of a new measure based on a translated version of the Psychological Consequences Questionnaire. *Submitted*

John Brodersen (MD, GP, PhD student) and Hanne Thorsen (MD, PhD, senior researcher) both Department and Research Unit of General Practice, University of Copenhagen, Denmark

This paper is dedicated to Professor Jill Cockburn with whom we have had rewarding discussions and who agreed with our project of adapting the PCQ for use especially in the setting of abnormal and false-positive screening mammography.

**Abstract**

**Objectives**
The aim of this study is to translate and adapt into Danish the original Australian version of the Psychological Consequences Questionnaire (PCQ), including an assessment of the content validity of the Danish version in the setting of abnormal screening mammography.

**Method**
The translation process included a bilingual panel and a lay panel. The adaptation and content validity process was ensured by conducting six focus group interviews including 34 women and fifteen individual telephone interviews.

**Results**
Six of the twelve original items in the PCQ were ambiguous or poorly worded and were reformulated increasing the number of items in the Danish version to 18 items. All of the 18 items were found relevant as negative consequences of abnormal screening mammography, however several areas were not covered. As a result, 15 new items were generated. Altogether these 33 items covered the negative consequences of abnormal screening mammography.

In the focus group interviews it was revealed that some issues could be raised both at invitation to screening, at time of screening, in the critical period, and after diagnosis. Other issues could be raised only after the women had been declared “free from” cancer suspicion.
Conclusion
The first part of the PCQ captures important negative consequences of abnormal screening mammography. However, it lacks content coverage for women after abnormal and false-positive screening mammography. A 33-item questionnaire was developed covering more negative consequences of an abnormal screening mammography than the original first part of the PCQ. The resulting 33-item questionnaire can be regarded as a new instrument that needs to be statistically tested for its’ psychometric properties.

Introduction
The distinction between normality and abnormality is not always clear cut and disease in populations exists as a continuum of severity rather than as an all-or-none phenomenon. Normality and abnormality are defined from different, often mutually exclusive, perspectives e.g. statistical, clinical, prognostic, and from the patients view. The prerequisite for medical screening is a technology capable of distinguishing normality from abnormality which can never be fully achieved.

Breast cancer is a substantial health problem in the western world. In Denmark the life time risk for breast cancer is 10% for women. When following the EU-recommended biennial breast cancer screening for women aged 50 - 69, the incidence of breast cancer diagnosed in each screening round is between 0.4 – 0.6%. In populations where a disease has low prevalence the positive predictive value of the technology used in screening procedures is low despite high specificity. A low positive predictive value results in a high rate of false positive results e.g. if a woman follows the EU-recommended screening programme for breast cancer for twenty years, her lifetime-risk for a false-positive screening mammography will be 20 - 25%, possibly even higher, and similar to other cancer screening programmes, 80 - 90% of abnormal screening mammographies are false-positive. False-positive screening mammography (hereafter referred to as a false positive) cause significant adverse economic consequences including costly follow-up tests. In addition, numerous studies have shown that women recalled for further investigations after an abnormal screening mammography, later confirmed as a false positive, experience significant adverse psychosocial effects such as worries, anxiety, sleeping problems, social isolation etc. The Psychological Consequences Questionnaire (PCQ) was developed by Jill Cockburn in Australia in 1992. The questionnaire is in two parts. The first measures the negative psychological consequences of the actual act of participation in breast cancer screening and the second measures
the positive consequences of participation. However, the PCQ has been applied mostly in studies of adverse consequences of false positives; in nine of ten studies, only the negative-consequences-part was used.

The items of the PCQ were developed after reviewing relevant questionnaires and interviewing women attending screening mammography, including women recalled for further investigation after an abnormal screening mammography.

McCaffery and Barrat underline the importance of high content validity when using questionnaires to measure the psychosocial outcome of screening. None of the studies using the PCQ to measure the consequences of false positives report examining the content relevance and content coverage before using it.

Women with the most negative psychosocial consequences of false positives are those who undergo surgery or early recall. However, these groups of women were not represented in the interviews preceding the development of the PCQ. Therefore, the original version of the PCQ may not capture all meaningful adverse psychosocial consequences of false positive; particularly longer-term or more severe consequences.

This paper reports the adaptation into Danish of the negative-consequences part of the PCQ, including a study to assess the content validity of the measure in the setting of abnormal and false-positive screening mammography.

Material and methods

Translation

Following an internationally accepted method, the translation process included a bilingual panel and a lay panel.

*Bilingual panel*. The first panel consisted of four members; all were female academics but without medical background. They were bilingual in English/Danish with Danish as their mother tongue. Each member of the panel independently translated the original Australian version of the PCQ into Danish. The panel members and the authors of this paper met and discussed the translations until they reached consensus. If consensus could not be reached, alternative translations of single items or sentences were left for the next panel. The bilingual panel produced the initial translations of the
items but it was left to the lay panel to decide upon the final wording and divide up the item if necessary.

Lay panel. The four members of the second panel were female state school teachers aged from 55-65; all taught Danish and none had ever taught English. For each item and each sentence of the instructions the lay panel was asked if the translated version of the PCQ was comprehensible and expressed in lay language. If it was not, panel members were asked to discuss and suggest alternative wording. Where the bilingual panel had suggested alternatives panel members were asked to choose the translation closest to lay Danish.

Adaptation and content validity in the setting of abnormal and false-positive screening mammography

Six focus group interviews were conducted with five to seven women in each group. The women had all been screened for breast cancer in the year prior to the interviews and all had experienced being told that their screening mammography was abnormal. All had also undergone additional medical procedures before the cancer suspicion was disproved. Women were grouped according to their additional examinations after abnormal screening mammography (table 1). Those living in a county offering breast cancer screening were included in the first four focus groups and those who had undergone opportunistic screening were included in the last two groups (table 1).

Table 1. All women in the focus groups had had an ultrasound examination and a clinical mammography (UCM) as a minimum after the abnormal screening mammography.

<table>
<thead>
<tr>
<th>No. of women</th>
<th>UCM</th>
<th>UCM and needle biopsy</th>
<th>UCM and surgical biopsy</th>
<th>UCM and Early recall</th>
<th>UCM and needle biopsy*</th>
<th>UCM and surgical biopsy*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age range (mean)</td>
<td>52 – 67 (58.2)</td>
<td>52– 67 (58.7)</td>
<td>52 – 60 (55.8)</td>
<td>52 – 69 (59.6)</td>
<td>28 – 71 (53.6)</td>
<td>49 – 65 (57.7)</td>
</tr>
</tbody>
</table>

*opportunistic screening

Issues for further exploration during the group interviews were decided upon after discussions with researchers, visits to a screening clinic, and after taking into account the relevant literature. The group interviews lasted approximately two hours and were divided into two parts. The first part was an open-ended discussion on consequences of an abnormal and a false positive screening mammography. The moderators also suggested more specific issues for discussion, i.e. physical, psychological, social and sexual consequences.
In the second part of the interview the women were asked to complete the Danish test version of the questionnaire; and comment on the wording of the items and the instructions, and easiness to complete. Timing was recorded. The women were then asked to “think aloud”\(^23\) and to discuss how they had responded to the questions according to the instructions. A cognitive group interview as an item-by-item discussion was also conducted.\(^24;25\) Finally, the women were asked if the items included in the test version covered all possible consequences they had experienced or were still experiencing after an abnormal and false positive screening mammography. The response options were tested for relevance and comprehensiveness, and the layout was tested for clarity.

If problems were revealed, the women in the focus groups were asked to suggest alternative questions, wording, or response options. If changes or new items were suggested in one group interview these were incorporated in the test version of the questionnaire presented to the following focus group.

Each interview was tape-recorded and independently audited by the two authors who conducted thematic analysis to determine key consequences of abnormal and false-positive screening mammography. Results were compared and if they did not correspond, the relevant sequences from the tape recording were re-audited and discussed until consensus was reached.

Ten additional telephone interviews were conducted with women at the time of their invitation for breast cancer screening and with five women from a county without screening. The women had received and completed the questionnaire before they were telephoned.

Results

*Translation*

The translations for five items were accepted by the lay panel without change. The translations for four other items were reformulated into more everyday language while still retaining the meaning of the original items. These changes were made because the lay panel found the language of the translated items too academic. The remaining three items (item 3, 4 and 5) created some problems. *Item 3:* Both panels considered this item to be ambiguous insofar as it appeared to be addressing two separate issues. The lay panel split item 3 into two because they regarded the meaning of the two words “unhappy” and “depressed” as covering different constructs.

*Item 4:* Item 4, asking if the women had been “scared” and “panicky”, was also found by both panels to be double-barrelled. In addition, the translator panel suggested two Danish synonyms for
the word “scared”. The lay panel split item 4 into two items because they regarded the meaning of
the two words “scared” and “panicky” to cover different constructs and chose one of the
alternatives suggested by the translator panel for the word ”scared”.
Item 5: This item was found to be double-barrelled by the bilingual panel and the translation into
Danish of the two words “nervous” and “strung up” was found redundant as they covered the same
construct. Therefore item 5 was kept as one item with only one Danish word covering the meaning
of “nervous” and “strung up”. The translation was accepted by the lay panel.
The changes in item 3, 4 and 5 resulted in a test version of the questionnaire containing 14 items.
The changes are summarised in table 2.

TABLE TWO ABOUT HERE

The bilingual panel suggested two different modes of response options. They argued that
linguistically in Danish some of the items were more correctly followed by a response option
expressing “frequency” instead of the uniform way of using “degree” in the original version. The
changes in response options were later accepted by the lay panel and the focus groups.

Adaptation and quality control of the translation

The introduction in the PCQ is: “Over the last week how often have you experienced the following
things because of thoughts and feelings about breast cancer”. The women in the lay panel and in
the focus groups argued that they did not have thoughts and feelings about breast cancer, they only
had thoughts. Therefore the word “feelings” was omitted in the Danish version.
The first focus group considered one of the translated items to be too academically formulated. This
item was reformulated and accepted by the next focus groups.

All items in the Danish version of the PCQ were found to be relevant by the participants in the
group interviews; however, three ambiguous and poorly worded items (item 1, 8 and 9) were
changed because they were misunderstood by the women. These three items were extended into
seven items, which cover in more detail the content of the original ones, see table 2.
The number of items therefore increased from 12 in the original Australian version of the PCQ to
18 items in the Danish version (PCQ-DK18), see table 2.
Table 2. In the instructions women were asked if they had been – had had – had felt – had found themselves or had experienced any of the following:

<table>
<thead>
<tr>
<th>The original Australian PCQ</th>
<th>The translated and adapted Danish version of the PCQ: – PCQ-DK18</th>
<th>The test version of a new questionnaire measuring negative consequences of abnormal screening mammography – PCQ-DK33</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) trouble sleeping (P)</td>
<td>taken me a long time to fall asleep</td>
<td>taken me a long time to fall asleep</td>
</tr>
<tr>
<td></td>
<td>slept badly</td>
<td>slept badly</td>
</tr>
<tr>
<td></td>
<td>taking sleeping tablets</td>
<td>taking sleeping tablets</td>
</tr>
<tr>
<td>2) change in appetite (P)</td>
<td>change in appetite</td>
<td>change in appetite</td>
</tr>
<tr>
<td>3) unhappy or depressed (E)</td>
<td>Unhappy</td>
<td>unhappy</td>
</tr>
<tr>
<td></td>
<td>Depressed</td>
<td>depressed</td>
</tr>
<tr>
<td>4) scared and panicky (E)</td>
<td>Scared</td>
<td>scared</td>
</tr>
<tr>
<td></td>
<td>Panicky</td>
<td>panicky</td>
</tr>
<tr>
<td>5) nervous or strung up (E)</td>
<td>Nervous</td>
<td>nervous</td>
</tr>
<tr>
<td>6) under strain (E)</td>
<td>under strain</td>
<td>under strain</td>
</tr>
<tr>
<td>7) keeping things from those who are close to you (S)</td>
<td>keeping things from those who are close to you</td>
<td>keeping things from those who are close to you</td>
</tr>
<tr>
<td>8) taking things out on other people (S)</td>
<td>Irritable</td>
<td>irritable</td>
</tr>
<tr>
<td></td>
<td>upset</td>
<td>upset</td>
</tr>
<tr>
<td>9) noticeable withdrawing from who are close to you (S)</td>
<td>withdrawn into myself</td>
<td>withdrawn into myself</td>
</tr>
<tr>
<td></td>
<td>quieter than normal</td>
<td>quieter than normal</td>
</tr>
<tr>
<td>10) difficulty doing normal things around the house (P)</td>
<td>difficulty doing normal things around the house</td>
<td>difficulty doing normal things around the house</td>
</tr>
<tr>
<td>11) difficulty meeting work or other commitments (P)</td>
<td>difficulty meeting work or other commitments</td>
<td>difficulty meeting work or other commitments</td>
</tr>
<tr>
<td>12) worried about your future (E)</td>
<td>worried about my future</td>
<td>worried about my future</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Response categories for the PCQ:
- not at all
- rarely
- some of the time
- quite a lot of the time

Domains: (E) Emotional, (P) Physical, (S) Social

The two items “less interest in sex” and “not felt like having my breast caressed” had an extra response category:
- not applicable

In the item “sick leave” the response categories were:
“0 days”, “1 – 2 days”, “3 – 4 days”, and “5 or more days”

*new item
Content validity

During the first group interview the women described their experiences as different in the critical period (from abnormal screening mammography until final false positive diagnosis) compared to the period after the final diagnosis. They argued that some issues could be raised only after they had been declared “free from” cancer suspicion. Other issues could be raised both at invitation to screening, at time of screening, in the critical period and, after diagnosis. They classified these issues as negative consequences of an abnormal screening mammography. Issues exclusively relevant to the time after final diagnosis were classified as long-term consequences of a false positive.

All questions in the PCQ-DK18 were found to be relevant as negative consequences of abnormal screening mammography. However, the women in the focus groups commented that several areas relevant to the negative consequences of abnormal screening mammography were not covered by the PCQ-DK18. As a result, 15 new items were generated; 12 by the first group and one new item by the second, third and fifth group each. Some of the 15 items covered new areas not covered by the PCQ namely, breast examination, sexuality and sick leave. Other new items covered areas such as anxiety, sense of dejection, symptoms and behavioural impact which were all related to the emotional and physical domains of the PCQ. The new items were validated in the subsequent groups. To verify that the content of the item suggested in fifth group was not relevant only for this particular group and for the sixth group, the tape recordings of the preceding four groups were re-audited. The content of the new items is listed in table 2:

The resulting questionnaire covering the negative consequences of abnormal screening-mammography contains 33 items (PCQ-DK33), see table 2.

The two modes of response options (“frequency” and “degree”) were accepted by the focus groups. In two new items concerning sexuality the interviewees wanted an extra response category; “not applicable”. For one new sick-leave-item the response categories: “0 days”, “1 – 2 days”, “3 – 4 days”, and “5 or more days” were accepted by the groups. The different response options are given in table 2.

The women participating in the telephone interviews all accepted the items in the PCQ-DK33 if given a thorough explanation on the front page about the purpose of the questionnaire.
Discussion

This study revealed that the adequacy and applicability of the PCQ in measuring psychosocial consequences of false positives is doubtful.

The translation and adaptation process for the present study utilised an internationally established and accepted method comprising two translation panels; a bilingual panel and a lay panel followed by field-testing for face and content validity via focus group interviews. An advantage of this method was that quality of the translation was checked and re-checked by bilingual translators and lay people and re-re-checked with sample respondents from the target population via the focus groups. This ensured a quality control during the translation and adaptation process. An a posteriori back translation which is often recommended was not used as this can be misleading and is not a guarantee for quality in the translation. However, four female state school teachers who formed the lay panel may have been too well-educated because the language of at least one item was regarded by the target population as being too academic. These pitfalls may not have been revealed if only a forward-backward translation procedure had been conducted.

Problematical items may undermine the validity of the data collected with a questionnaire. For example, if an item contains two ideas, some respondents will answer “yes” only if both parts are true: however, others will answer “yes” if either part is true. The final result may not reflect the actual state of affairs. If an item is poorly worded, it may be misunderstood. The translation and adaptation process showed that several items in the original version of the PCQ were ambiguous or poorly worded. Women taking sleeping tablets or having problems falling asleep would not know how to respond to item 1 in the PCQ “Had trouble sleeping”. This lack of detail about sleeping problems resulted in three additional items. Splitting items containing more than one idea and reformulating unclear items increased the total number in the questionnaire.

One of the results from the focus group discussions on content validity confirmed the findings by Cockburn when developing the PCQ: women having an abnormal screening mammography did not find items asking about positive consequences of participation in breast screening relevant before final diagnosis. The women in the focus groups argued that they only experienced negative consequences of having an abnormal screening mammography. Therefore, only the items in the first part of the PCQ that measures the negative psychological consequences of participation in breast cancer screening were relevant for the women in the critical period from abnormal screening until final diagnosis.
Furthermore, the second part of the PCQ that measures the positive consequences of the actual act of participation in breast screening was not relevant at all for the women having a false positive. The women argued that after being declared “free from” cancer suspension it was neither negative nor positive psychosocial consequences they had experienced. The impact of a false positive resulted in changes to predominantly one area such as, relation to social network or existential values. These comments suggest that a totally new questionnaire would be required to cover the long-term psychosocial consequences of a false positive. It was decided that the development and validation of such a new measure would constitute a separate study.

The study highlighted the importance of including the target population when scrutinising the content validity of an instrument. All women participating in the focus groups had experienced an abnormal screening mammography, which later was confirmed to be false-positive. Focus group participants were selected according to the additional examinations after abnormal screening mammography. The advantage of this homogeneity was that the discussion in the group was focused on experiences they had all had instead of having a discussion where the women compared different experiences between one another. As a result 15 new items were generated to cover the lack of content coverage of the original PCQ in the setting of abnormal screening mammography.

The women interviewed at the recall clinic during the development of the original PCQ did not know if they had breast cancer or not. To conduct interviews with women under threat of breast cancer and therefore under stress may not be ideal. In the present study all interviewees knew that their abnormal screening mammography was false. This may have influenced their willingness to discuss their crises in the critical period. Studies have shown that women undergoing clinical mammography only in addition to screening mammography are those with less negative psychosocial consequences compared to those undergoing needle biopsy, surgery or early recall. Therefore, it is worth mentioning that 12 of the 15 new items were generated during the first focus group interview with participants who had undergone the least distressing additional examinations.

There is no doubt that the PCQ captures important negative consequences of abnormal screening mammography. However, scrutinizing the content validity in this setting has shown that the PCQ does not cover all consequences for women after abnormal screening mammography. On the other hand, developing 15 new items only ensured a higher content validity and is no guarantee for perfect content coverage. However, the fact that only three new items were generated in the second to fifth focus group indicated a high degree of data saturation.
Conclusion
Testing the PCQ linguistically and for content validity in a setting of abnormal screening mammography until data saturation has resulted in a new questionnaire with 33 items instead of the original twelve items. This extended questionnaire covers more negative consequences of abnormal screening mammography than the original Australian version. However, it does not cover all possible long-term consequences of false positives. Before using the new instrument a validation study should be conducted to assess its psychometric properties.

Acknowledgement
We wish to thank the staff at the breast cancer screening clinic and the recall clinic at the Copenhagen University Hospital and at the Surgical Department of Hoersholm Hospital for recruiting women to this study.
Reference List


8. Discussion of article 2

8.1. Major findings

The psychosocial consequences experienced in the critical period from abnormal screening mammography until final false-positive diagnosis differ in some respects from those experienced after having the final diagnosis. In this critical period women do not experience any positive psychosocial consequences, only negative consequences. The negative part of the Psychological Consequences Questionnaire (PCQ) (including 12 items) had six problematic items. Three of the items were found to be double-barrelled in meaning; that is, they asked two questions in one item. Three other items were ambiguous and could be misunderstood. Translating and adapting the negative part of the PCQ into Danish and reformulating the six problematic items into twelve non-ambiguous items resulted in an 18-item Danish version of the PCQ. These 18 items were all relevant for women having an abnormal screening result; however, the 18 items did not cover the large spectrum of negative psychosocial consequences experienced by the women. To cover these consequences 15 additional items were developed, resulting in a new 33-item questionnaire.

8.2. Assessment of methods

8.2.1. Translation and adaptation

It is not straightforward to translate a questionnaire from the original language and culture into another. Several pitfalls should be taken into account.

1. Languages are changing over time and thereby the meaning of words and phrases. For example, the word “gay” has changed dramatically over the recent decades.

2. Cultural differences may also create problems. For example, a bag of sugar comes in different sizes of weight in different countries so asking “can you lift a bag of sugar?” is inconsistent from one country to another.

3. A too literate translation may change the meaning of an item. A literal translation of the PCQ-item “I have been panicky” into Danish is “Jeg har været ved at gå i panik”. The meaning of “panicky” in English is close to being scared while the meaning of “panik” in Danish is close to being terrified. Another example from the Nottingham Health Profile (NHP) is the translation of the item “I’m waking up in the early hours of the morning” which in English means that you wake up at 1-2 a.m. but in Danish means that you wake up at 5-6 a.m.

4. Problems occur especially when metaphors or items formulated in jargon are translated because they are specific to a culture at a specific time age. For example, the item “I’m feeling on edge” is taken from the NHP that was developed in the United Kingdom in the 1980s. This expression is nowadays obsolete in English and has been exchanged with “I’m feeling edgy”. None of these two expressions make sense if translated literally into Danish.

Therefore, questionnaires used over many years should be checked frequently to ensure that the meaning of the items and the instructions remain relevant. If it is necessary during the translation and adaptation process to change an item and select more specific or simpler words, the meaning of the original item should be maintained.

In paper 2 the translation and adaptation process of the PCQ was conducted using a bilingual panel and a lay panel. In the bilingual panel the discussion of the pitfalls mentioned above was taken into account and discussed. As a result of the discussion consensus could not always be reached and the
bilingual panel suggested two or three alternatives for such items. It was then for the lay panel to decide which of the alternatives was closest to lay-Danish. For every item and instruction the lay panel was asked
1. What do you think is the meaning of this question?
2. Can you answer the question?
3. Do you think it is necessary to rephrase this question? And if yes, how?
These questions were also suggested to ensure the quality in the translation process. For detailed information of the translation and adaptation process see appendix III (page 123).

8.2.2. Assessment of content validity
Before the focus group interviews were conducted, relevant literature was searched. There were two purposes for conducting this literature search: writing an interview-guide and reviewing the adequacy of measurement of short-term and long-term consequences of false-positive screening mammography (hereafter referred to as a false positive). Individual interviews with three women who had had a false positive were conducted to find important topics missing from the literature; issues about death, thoughts about funeral arrangements and writing a personal will were revealed. A well established method to develop new items, scales and questionnaires is to interview members of a target population either by conducting single interviews or focus group interviews. The target population is the optimal source of information on their conditions and experiences. In the present study, focus group interviews with women having a false positive were used to assess content validity. The women were grouped according to their additional examinations after abnormal screening mammography to avoid discussion on different examination procedures in the critical period. In the group interviews the moderator emphasised that all participants had had an abnormal screening mammography and, for example, a needle biopsy examination before the final diagnosis. The women could then focus on the feelings, thoughts and consequences they had experienced after the screening mammography.

The group interviews began with an open-ended discussion to capture issues not described in the literature and not found in the pilot interviews. As a result of the open-ended discussions “sick-leave” and “feeling less attractive” were revealed as new issues. Some women had been on sick-leave in the critical period. In the groups including women who had had surgery the women were unhappy about the scar on their breast - especially if the skin surrounding the scar had an indentation. The women felt less attractive, both in a sexual and social context. For example, when undressing in a public swimming pool they felt less attractive compared to women with no breast deformities. The women also discussed that they would feel less attractive after a mastectomy. After the open-ended discussion the moderator introduced specific issues to be discussed in the groups. When discussing the issues about death, thoughts about funeral arrangements and writing a personal will the groups agreed that these issues would be unacceptable in a questionnaire in the setting of screening mammography. Therefore, items covering these issues were not included. However, the item “I have kept busy to take my mind off things” was found acceptable and non-offensive. This item covers issues related to thoughts about death. In one of the group interviews a woman told the group that she had cleared up the whole house to push away her thoughts about death. Another had selected a burial plot with her husband. When issues about sexuality were discussed in the first group interview, none of the participants confirmed that they had experienced a change in their sexuality. Answers to speculations on why these women did not raise the topic of sexuality could be:
• this group of women did not experience any negative impact on their sexuality
• the women found that a discussion about sexuality was taboo among unfamiliar women
• the presence of a male interviewer was a hindrance

Both authors felt that the women were too embarrassed discussing sexuality in the group. Therefore, in the next focus group it was decided that the male author should leave the interview while discussing sexuality. The female author then began a discussion about sexuality in general and turned the discussion into psychosexual consequences of screening mammography resulting in topics generating two new items. The topic of sexuality and the two sexuality-items were discussed in the following four group interviews with the male moderator participating and were found acceptable and not offensive.

The PLISSIT model offers an approach to communicating with people about sexuality. The “P” in the model stands for “Permission” and is meant as a reassurance given by the interviewer to people that they are not abnormal or perverted and it is also a permission to talk about sexuality. In the second focus group the female author gave permission to talk about sexuality and in the following four group interviews the two sexuality items were the permission to talk about sexuality.106

8.3. Justification of conclusion

Six of the original negative PCQ-items were found ambiguous which might have disturbed the sensitivity of the original version of the PCQ. However, the 18 items in the Danish version of the negative part of the PCQ were all found to be relevant in the group interviews. Therefore, the negative part of the PCQ might capture some of the psychosocial consequences women are experiencing after an abnormal screening mammography.

To cover more of the spectrum of the psychosocial consequences of abnormal screening mammography, 15 items were added to the Danish 18-item version of the measure. Some of these negative consequences covered by the 33 items may last after final diagnosis. However, the 33 items were not equal to the specific consequences occurring after final diagnosis.

The 33-item questionnaire can be regarded as a new instrument, because the Danish version of the PCQ has not been statistically validated and 15 new items have been added. Therefore, the 33-item questionnaire ought to be statistically tested for its psychometric properties before it is used.

8.4. Contribution to the current knowledge

Article 1 showed that surveys using the PCQ all consistently reported negative short-term psychosocial consequences of false positives.104 However, the results were ambiguous in surveys on long-term consequences. A reason for this inconsistency could be the lack of content coverage of the negative part of the PCQ:

• in terms of negative psychosocial consequences of abnormal screening mammography
• and measuring specific psychosocial consequences only relevant after final diagnosis

The present study supports the conclusion of article 1: any conclusions about the long-term psychosocial consequences of false positives must remain tentative.104

In two studies pain in the breast and reduced sexual sensitivity of the breast after surgery have been reported as consequences of false positives.107,108 These issues are not covered by the PCQ.52 In the present study the item, “felt less attractive” partly covers these issues.

Increased frequency of breast self-examination has been reported as a consequence of false positives.62,99-101 No items in the PCQ ask about breast self-examination. In the present study two new items; “I have examined my breasts” and “I have examined my breasts in the mirror” were generated to cover the topic.
9. Article 3

Validation of a condition-specific measure for women having an abnormal screening mammography. Submitted

John Brodersen (MD, GP, PhD student)*, Hanne Thorsen (MD, PhD, senior researcher)* and Svend Kreiner (MSC, associate professor)**.
*Department and Research Unit of General Practice, ***Department of Biostatistics
All three: Institute of Public Health, University of Copenhagen, Denmark

Abstract
Objectives
The aim of this study is to assess the validity of a new condition-specific instrument measuring psychosocial consequences of abnormal screening mammography (PCQ-DK33).

Material and method
The draft version of the PCQ-DK33 was completed on two occasions by 184 women who had received an abnormal screening mammography and on one occasion by 240 women who had received a normal screening result. Item Response Theories and Classical Test Theories were used to analyse data. Construct validity, concurrent validity, known group validity, objectivity and reliability were established by item analysis examining the fit between item responses and Rasch models.

Results
Six dimensions covering: anxiety, behavioural impact, sense of dejection, impact on sleep, breast examination and sexuality were identified. One item belonging to the dejection dimension had uniform differential item functioning. Two items not fitting the Rasch models were retained because of high face validity. A sick-leave item added useful information when measuring side effects and socioeconomic consequences of breast cancer screening. Five “poor items” were identified and should be deleted from the final instrument.

Conclusion
Preliminary evidence for a valid and reliable condition-specific measure for women having an abnormal screening mammography was established. The measure includes 27 “good” items measuring different attributes of the same overall latent structure – the psychosocial consequences of abnormal screening mammography.
Introduction

If a woman follows the EU-recommended biannual breast cancer screening programme for twenty years her lifetime-risk for a false-positive screening mammography will be 20 - 25%, possibly even higher.¹ In the UK, where three-yearly screening is offered to women aged 50 - 65, more than 50,000 women per year will receive a false-positive screening mammography (hereafter referred to as a false positive).² False positives cause significant adverse consequences including costly follow-up tests and an increased use of health care services.³-⁵ In addition, numerous studies have shown that women recalled for further investigations after an abnormal screening mammography, later confirmed as a false positive, experience significant adverse psychosocial effects.⁶ In these studies a variety of questionnaires have been used to measure the adverse effects. Except for one questionnaire most of the instruments have been developed for other purposes and have a more or less generic character.⁷

A requirement for the validity of a questionnaire is that it has high content relevance and high content coverage.⁸ A condition-specific measure insures higher content coverage compared to measures developed for generic conditions.⁹ A review has shown that some of the most frequently used generic measures in the setting of breast cancer screening: the General Health Questionnaire, the Hospital Anxiety and Depression Scale and the State-Trait Anxiety Inventory, have problems with language, content relevance and content coverage.⁷ McCaffery and Barrat underline the importance of high content validity when using questionnaires to measure psychosocial consequences of screening.¹⁰

The Psychological Consequences Questionnaire (PCQ) was developed in 1992 by Jill Cockburn to measure the short-term psychosocial consequences of the actual act of participation in breast cancer screening. It consists of twelve items covering negative aspects and ten items covering positive aspects of participation.¹¹ The full PCQ (negative and positive items) has been used in only one study of the process of participating in breast cancer screening.¹² The negative items have predominantly been used in studies of the adverse consequences of false positive and not of the consequences of the actual participation.⁶ However, the content validity of the PCQ has never been tested in the setting of abnormal and false positive screening mammography.⁷

When summing raw scores of items in a scale an assumption of unidimensionality is made. That is that the items describe different aspects of the same construct and can be added.¹³;¹⁴ When the response options are categorical on an ordinal scale as in many questionnaires Item Response Theory (IRT) and Rasch models can be used to assess the psychometric properties of the
questionnaire. The Rasch models provide formal representation of perfect measurement. Where items are shown to fit a Rasch model the measure can be shown to posses criterion-related construct validity, to be objective, sufficient and, therefore, also reliable. If a measure has specific objectivity comparison between two persons with one part of a scale does not differ systematically from comparisons using another set of valid items from the scale. Similarly, comparison of two persons should not depend on measurements on other persons. For these reasons the Rasch model is considered a valuable “gold standard” against which measures should be compared. Reliability and different aspects of validity can also be assessed using Classical Test Theory (CTT). The relation between CTT and IRT has been described by Holland and Hoskens and it may be an advantage to combine the two theories. Item analyses by Rasch models explore in depth the degree to which the requirements of construct validity are met. Items are assumed to monotonically relate to one dimension and they are assumed to be locally independent. It is also assumed that there is no differential item functioning (DIF) that is, where an item functions differently in subpopulations such as in an intervention group and a control group. The sufficiency of the model support computation of the raw scores. IRT analyses also explore how the items included in each dimension are interrelated and ordered on a latent trait (e.g. psychosocial consequences of a false positive). The strength of analyses based on the Rasch Model is that the model is build on pre-assumptions closer to reality than analyses based on Classical Test Theory. The Rasch models describe how item responses depend both on person and item parameters. The person parameter is assumed to be unidimensional but item parameters may be multidimensional when item responses are ordinal categories.

The purpose of this study was to validate a new condition-specific instrument measuring psychosocial consequences of abnormal screening mammography (PCQ-DK33) using both IRT and CTT.

Material and methods
A qualitative study to assess the content validity of the PCQ was conducted in a setting of abnormal screening mammography. The qualitative study highlighted the need to make radical changes to the questionnaire if it was to be used in this setting. Therefore, the draft version of the questionnaire statistically tested in the present study can be regarded as a new condition-specific instrument with 33 items (PCQ-DK33, see the summary in table 4).
Data were collected at two screening centres: the Copenhagen University Hospital and Odense University Hospital.

**Group 1 - Time I.** Over a period of 20 weeks from November 2002 all women who were recalled because of an abnormal screening mammography were consecutively included in the study. Before any further examinations to establish if the abnormal screening result was true or false the women were asked to complete two questionnaires: 1) the PCQ-DK33 with the items randomly ordered and 2) the Danish version of the Nottingham Health Profile (NHP).

The NHP is a questionnaire measuring health status. It was originally developed in the UK and has been adapted into a large number of languages including Danish. The NHP consists of six sections covering energy, pain, emotional reaction, sleep, social isolation and physical mobility. It was selected as a comparator to assess concurrent validity for the present study because emotional reactions and sleep problems are well covered by the measure. It was hypothesised that the new instrument measuring psychosocial consequences of abnormal screening mammography would converge with the emotional section of the NHP and diverge from the pain and mobility sections.

**Group 1 - Time II.** Two weeks after having completed the first questionnaire package the women were sent the draft version of the PCQ-DK33. At this point most of the women would know if their abnormal screening result was false positive or if they had breast cancer. They were asked to complete the questionnaire and return it in an enclosed stamped addressed envelope. It was hypothesised that women diagnosed with breast cancer experienced more severe psychosocial consequences than women with a known false-positive screening result (known group validity and responsiveness). It was also hypothesised that there would be a decrease in the negative psychosocial consequences from abnormal to known false-positive screening result.

**Group 2.** For each woman in Group 1 who had completed the questionnaires at Time I another two women were sent the draft version of the PCQ-DK33 and asked to return the completed questionnaire in an enclosed stamped addressed envelope. These women had had a normal screening mammography at the same time and at the same clinic as the women in Group 1, Time I. The recruitment procedure for women in Group 2 was: if the screening mammography of a woman was abnormal then the two women having a normal screening mammography and being screened just before and after this woman were included in Group 2. In contrast to Group 1, where informed consent was obtained at the recall clinics, it was not possible to obtained informed consent before posting the test-questionnaire to the women in Group 2. For ethical and legal reasons the test-questionnaire was posted to the women by the screening clinics anonymously and only the age of
the participants was disclosed for the researchers. Therefore, it was not possible to send reminders to the women in Group 2. It was hypothesised that women with an abnormal screening result experienced more severe psychosocial consequences than women with a normal screening result (known group validity).8,25

In Group 1 – Time I, 220 women were eligible. Of these, 16 (7.3%) were not invited to participate due to sick leave among the clinic staff. Of those asked to participate 90.2% agreed to complete the questionnaires. At Time II, 90.8% returned the PCQ-DK33 after one reminder. In Group 2, 400 women received the questionnaire by post and 60% were returned. There were no statistically significant differences in mean and range of age between the women in the three groups. Figure 1 illustrates the data collection including the numbers of women in each subgroup.

Figure 1. Data collection of the PCQ-DK33

<table>
<thead>
<tr>
<th>Group 1, Time I</th>
<th>Group 1, Time II</th>
</tr>
</thead>
<tbody>
<tr>
<td>184 women with abnormal screening mammography</td>
<td>167 of 184 women with:</td>
</tr>
<tr>
<td></td>
<td>• 89 false-positive</td>
</tr>
<tr>
<td></td>
<td>• 61 breast cancer</td>
</tr>
<tr>
<td></td>
<td>• 17 undiagnosed</td>
</tr>
</tbody>
</table>

Three questionnaires were returned without being completed. Among the remaining 588 questionnaires 0.3% - 1.9% randomly distributed missing values per item were observed. Item responses were analysed by the conditional distribution of items given total person scores in order to avoid assumptions on the distribution of the latent trait being measured. The pairwise estimation procedure implemented in software program RUMM2020 was used to estimate the item parameters.26,27 The analysis of the fit of item responses to the Rasch model were based on analyses of residuals comparing observed to expected item responses, both for separate individuals and for different score groups. The overall fit of the model was assessed by the Wright-Panchapakesan $\chi^2$ statistic summarising standardised residuals over score groups and items.28 Item fit statistics summarising standardised residuals in different score groups were used to identify misfitting items. Data from Group 1, Time I and Time II and data from Group 2 were pooled for IRT analyses. However, the sick-leave item was not included in the Rasch analyses because the response options...
differed entirely from the response options in the remaining 32 items. Differential item functioning relative to person covariates was checked by analyses of variance examining the degree to which individual residuals for specific items depended on the covariates. Absence of evidence of interaction between the covariates and the estimated trait parameters were taken as evidence of DIF being uniform. A small subgroup of 17 women undiagnosed two weeks after their abnormal screening mammography was not included in the analyses of DIF. Finally, the assumption of local independence was checked by examination of the degree to which individual residuals were correlated. Reliability was assessed by Cronbach’s alpha defining the lower bound for the test-retest correlation of the raw scores \(^{29,30}\) and by the so-called Person Separation Index calculating the lower bound for the test-retest correlation of the estimated values of the latent trait being measured.\(^{31}\) Data were analysed with CTT by using the software SPSS for Windows version 13.0 and the software Mplus 2.14 for confirmatory factor analysis.\(^{32}\) For IRT analyses the software RUMM2020 was used.\(^ {26}\) The study was approved by the local ethical committee.

Results
The initial item analysis of the complete set of 32 items rejected the Rasch model. Strong evidence of local dependence indicated that the PCQ-DK33 was not unidimensional. The subsequent separate analyses confirmed the multidimensionality expected from the qualitative study on face and content validity of the instrument.
Six items covering anxiety formed one dimension and none of these items had DIF in either of the subgroups. In one of the six items, “felt terrified” (no 29), the thresholds of the response categories were not in order (figure 2). Of eight items covering the impact on behaviour after abnormal screening mammography, seven items fitted the Rasch model and no DIF was observed. Among the seven behavioural items the thresholds of item “difficulty doing everyday things around the house” (no 28) were not in order (figure 3). Six items describing the sense of dejection and sadness after abnormal screening mammography fitted the Rasch model forming one dimension. However, the item “felt sad” (no14) had uniform DIF in two of four subgroups as shown in figure 4. After deleting this item the five remaining items still fitted the Rasch model with all thresholds in order and no DIF.
Figure 2. Category Probability Curve of item 29 “terrified”

The logit scale from -2 to +8 on the x-axis symbolises the latent trait of anxiety, with the severity of anxiety increasing towards the right. The y-axis symbolises the probability of affirming the response categories: 0 “not at all”, 1 “a bit”, 2 “quite a bit” and 3 “a lot”.

Figure 3. Category Probability Curve of item 28 “difficulty doing things around the house”

The logit scale from -3 to +4 on the x-axis symbolises the latent trait of behavioural impact, with the severity of negative impact increasing towards the right. The y-axis symbolises the probability of affirming the response categories: 0 “not at all”, 1 “a bit”, 2 “quite a bit” and 3 “a lot”.
Figure 4. Item characteristics curves (ICC) of item 14 “felt sad” showing uniform differential item functioning (no statistical significant difference between the slopes of the three ICC) between the group of women diagnosed with breast cancer (i14bc), the group of women having normal screening mammography (i14no) and the remaining women – women having abnormal and a false positive screening mammography (i14re).

The logit scale from -7 to +5 on the x-axis symbolises the latent trait of sense of dejection, with the severity of dejection increasing towards the right. The y-axis symbolises the values of responses options: 0 “not at all”, 1 “a bit”, 2 “quite a bit” and 3 “a lot”. Locn symbolises the item location.

From a content perspective the Rasch analyses confirmed three more dimensions each with two items. These three dimensions described impact on breast examination, sleep and sexuality. However, in the breast examination dimension both items had uniform DIF. The item “examined my breasts” (no 16) had uniform DIF in the group diagnosed with breast cancer compared to the remaining groups (figure 5) and the item “examined my breasts in the mirror” (no 21) had uniform DIF in the group with normal screening mammography compared to the remaining groups (figure 6). In the sexuality dimension the thresholds of the item “less interest in sex” (no 31) were not in order (figure 7).
Figure 5. Item characteristics curves (ICC) of item 16 “examined my breasts” showing uniform differential item function (no statistical significant difference between the slopes of the two ICC) between the group of women diagnosed with breast cancer (i16bc) and the remaining women – women having an abnormal, a false positive and a normal screening mammography (i16re).

The logit scale from -6 to +4 on the x-axis symbolises the latent trait of breast examination, with the severity of breast examination increasing towards the right. The y-axis symbolises the values of responses options: 0 “not at all”, 1 “a bit”, 2 “quite a bit” and 3 “a lot”. Locn symbolises the item location.
Figure 6. Item characteristics curves (ICC) of item 21 “examined my breasts in the mirror” showing uniform differential item functioning (no statistical significant difference between the slopes of the three ICC) between the group of women with a normal screening mammography (i21no) and the remaining women – women having an abnormal, a false positive and true positive (breast cancer) screening mammography (i21re).

The logit scale from -6 to +4 on the x-axis symbolises the latent trait of breast examination, with the severity of breast examination increasing towards the right. The y-axis symbolises the values of responses options: 0 “not at all”, 1 “a bit”, 2 “quite a bit” and 3 “a lot”. Locn symbolises the item location.
Figure 7. Category Probability Curve of item 31 “not felt like having my breast caressed”

The logit scale from -5 to +4 on the x-axis symbolises the latent trait of sexuality, with the severity of negative impact of sexuality increasing towards the right. The y-axis symbolises the probability of affirming the response categories: 0 “not at all”, 1 “a bit”, 2 “quite a bit” and 3 “a lot”.

Besides testing the six identified dimensions for differential item function in the sampled groups (figure 1), DIF was also tested for age and screening centre. These tests showed no DIF. There was no local dependency among the items in the identified dimensions.

The two sleeping items fitting the Rasch model and the item “taking sleeping tablets” are from a content point of view equal to three of the five items included in the sleep section of the NHP. Rasch analyses of the sleep section of the NHP showed that four of the five sleeping items fitted the Rasch model except for the sleeping-tablet item.

The Wright-Panchapakesan Chi Squared fit Statistics, the Person Separation Index and the Cronbach’s alpha of the six dimensions fitting the Rasch model in the PCQ-DK33 are listed in Table 1. The fit statistic of the behavioural subscale is marginally significant (p = 0.039). Adjusting p-values in order to control the false discovery rate and so avoid spurious significant results due to multiple testing suggested that the result should be regarded as insignificant.33
Table 1. Wright-Panchapakesan (WP) Fit statistics, Person Separation Index and the Cronbach’s alpha of six dimensions in the PCQ-DK33

<table>
<thead>
<tr>
<th>Dimensions</th>
<th>WP $\chi^2$</th>
<th>Degrees of freedom</th>
<th>p</th>
<th>Person Separation Index</th>
<th>Cronbach’s alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anxiety (6)</td>
<td>61.81</td>
<td>52</td>
<td>0.166</td>
<td>0.94</td>
<td>0.92</td>
</tr>
<tr>
<td>Behavioural (7)</td>
<td>67.71</td>
<td>49</td>
<td>0.039</td>
<td>0.88</td>
<td>0.86</td>
</tr>
<tr>
<td>Sense of dejection (5)</td>
<td>55.14</td>
<td>40</td>
<td>0.056</td>
<td>0.93</td>
<td>0.89</td>
</tr>
<tr>
<td>Sleep (2)</td>
<td>8.49</td>
<td>9</td>
<td>0.486</td>
<td>0.89</td>
<td>0.90</td>
</tr>
<tr>
<td>Breast examination (2 or 4*)</td>
<td>19.60(28.34*)</td>
<td>8(21*)</td>
<td>0.01(0.131*)</td>
<td>0.68(0.70*)</td>
<td>0.71</td>
</tr>
<tr>
<td>Sexuality (2)</td>
<td>7.06</td>
<td>10</td>
<td>0.720</td>
<td>0.81</td>
<td>0.83</td>
</tr>
</tbody>
</table>

*after item split according to the uniform DIF found.

For each subgroup table 2 shows the mean score, the standard error of mean and the standard deviation for all six dimensions fitting the Rasch model.
### Table 2. Mean scores, standard error of mean, and standard deviation of all six dimensions

<table>
<thead>
<tr>
<th>Dimensions</th>
<th>Anxiety</th>
<th>Behavioural</th>
<th>Dejection</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>B</td>
<td>C</td>
</tr>
<tr>
<td>No**</td>
<td>178</td>
<td>52</td>
<td>97</td>
</tr>
<tr>
<td>Mean</td>
<td>6.39</td>
<td>7.38</td>
<td>3.04</td>
</tr>
<tr>
<td>S.E. Mean</td>
<td>0.36</td>
<td>0.73</td>
<td>0.42</td>
</tr>
<tr>
<td>Std. deviation</td>
<td>4.77</td>
<td>5.30</td>
<td>4.18</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dimensions</th>
<th>Sleep</th>
<th>Breast examination</th>
<th>Sexuality</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>B</td>
<td>C</td>
</tr>
<tr>
<td>No**</td>
<td>179</td>
<td>52</td>
<td>97</td>
</tr>
<tr>
<td>Mean</td>
<td>2.10</td>
<td>2.48</td>
<td>1.10</td>
</tr>
<tr>
<td>S.E. Mean</td>
<td>0.15</td>
<td>0.34</td>
<td>0.18</td>
</tr>
<tr>
<td>Std. deviation</td>
<td>2.00</td>
<td>2.43</td>
<td>1.73</td>
</tr>
</tbody>
</table>

A) Group 1 - Time I B) Group 1 - Time II, women diagnosed with breast cancer. C) Group 1 - Time II, women with known false positive D) Group 1 - Time II, women undiagnosed. E) Group 2

**Number of subjects included in the Rasch analyses. *Mean scores are estimated from the means of the person locations on the latent trait according to the uniform DIF found in item 16 & 21.

# The Standard Errors of Means (S.E. mean) are estimated from the S.E. means of the person locations.

Items numbered 1, 7, 9, 10, 25, 27 and 30 did not fit the Rasch model. The face validity of these items was checked by re-auditing tape recordings from the focus group interviews conducted during the adaptation of the PCQ into Danish. The item “less attractive” (no 1) and the item “busy to take mind off things” (no 10) had significant face validity for women who had had surgery or had been on early recall after abnormal screening mammography.

Cronbach’s alpha increased when deleting items 7, 25, 27 and 30 and dropped when deleting items 1, 9 and 10. This indicated that items 7, 25, 27 and 30 were “poor” items because alpha is expected to decrease when valid items are deleted from a summated score.

A subsequent confirmatory factor analysis was conducted to estimate how the six dimensions and the three single items (no’s 1, 10 and 14) were correlated. The analysis revealed a positive correlation between all six dimensions and the three single items. Only very weak evidence was disclosed against a model assuming that one latent trait lies behind the six dimensions and the three single items of the PCQ-DK33 (p = 0.0463).
When testing for *known group validity* a statistically significant difference was found between women having an abnormal screening mammography and those having a normal screening mammography. In all six dimensions and in the three single items the p-value of the Pearson Chi-Square was less than 0.0005. In the single item about sick leave (no 33) the p-value was 0.026. There was a statistically significant difference between women diagnosed with breast cancer and those having a known false positive in five of the six dimensions and in the three single items (no’s 1, 10 and 33) with the highest p-value as 0.016. The sexuality subscale could not differ between women diagnosed with breast cancer and those having a known false-positive screening result. Additional analyses on the sexuality subscales found no statistically significant difference between women having an abnormal screening mammography and those diagnosed with breast cancer. They also showed no difference between women having an abnormal screening mammography and those with a known false positive.

As a test of *concurrent validity* the Pearson Correlation was established between the sections of the NHP and all the 27 “good” items (24 items in the six identified dimensions and items 1, 10 and 14). The Pearson Correlation was also established for each of the six Rasch-fitting dimensions. The results confirmed the hypothesis made before the analysis (table 3).

Table 3. Concurrent validity (convergent and divergent validity) of 27 “good” items and the six identified dimensions of the PCQ-DK33 and the sections of the NHP at Time I - Group 1.

<table>
<thead>
<tr>
<th>NHP sections</th>
<th>27 “good” items</th>
<th>Anxiety</th>
<th>Behavioural</th>
<th>Dejection</th>
<th>Sleep</th>
<th>Breast examination</th>
<th>Sexuality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Energy level</td>
<td>0.43</td>
<td>0.30</td>
<td>0.49</td>
<td>0.37</td>
<td>0.30</td>
<td>-0.05</td>
<td>0.30</td>
</tr>
<tr>
<td>Pain</td>
<td>0.15</td>
<td>0.15</td>
<td>0.14</td>
<td>0.13</td>
<td>0.14</td>
<td>-0.13</td>
<td>0.09</td>
</tr>
<tr>
<td>Emotional reaction</td>
<td>0.70</td>
<td>0.65</td>
<td>0.68</td>
<td>0.67</td>
<td>0.55</td>
<td>0.15</td>
<td>0.35</td>
</tr>
<tr>
<td>Sleep</td>
<td>0.48</td>
<td>0.38</td>
<td>0.40</td>
<td>0.39</td>
<td>0.69</td>
<td>0.12</td>
<td>0.20</td>
</tr>
<tr>
<td>Social isolation</td>
<td>0.26</td>
<td>0.27</td>
<td>0.29</td>
<td>0.24</td>
<td>0.25</td>
<td>0.01</td>
<td>0.11</td>
</tr>
<tr>
<td>Physical mobility</td>
<td>0.16</td>
<td>0.10</td>
<td>0.17</td>
<td>0.10</td>
<td>0.10</td>
<td>-0.17</td>
<td>0.13</td>
</tr>
</tbody>
</table>

The correlations are calculated as coefficients from the Pearson Correlation.

A summary of the results from the psychometric analyses of the PCQ-DK33 are given in table 4.
Table 4. Summary of result from the psychometric analyses of the PCQ-DK33.

<table>
<thead>
<tr>
<th>The items of the PCQ-DK33 in order of appearance in the draft version</th>
<th>Subscales. Dimensionality and misfit to the Rasch model</th>
<th>Probability of fit to the Rasch model</th>
<th>Degrees of freedom</th>
<th>Chi Square</th>
<th>Cronbach’s alpha</th>
<th>Face validity</th>
<th>Single or “poor” item</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. less attractive</td>
<td>Misfit</td>
<td>0.093</td>
<td>7</td>
<td>12.241</td>
<td>Dropped</td>
<td>High</td>
<td>Single item</td>
</tr>
<tr>
<td>2. worried</td>
<td>Dejection</td>
<td>0.471</td>
<td>8</td>
<td>7.627</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. worried about my future</td>
<td>Anxiety</td>
<td>0.280</td>
<td>9</td>
<td>10.931</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. scared</td>
<td>Anxiety</td>
<td>0.208</td>
<td>9</td>
<td>12.098</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. irritable</td>
<td>Behavioural</td>
<td>0.513</td>
<td>7</td>
<td>6.236</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. quieter than normal</td>
<td>Behavioural</td>
<td>0.902</td>
<td>7</td>
<td>2.805</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. keeping things from those who are close to you</td>
<td>Misfit</td>
<td>&lt; 0.0005</td>
<td>8</td>
<td>50.659</td>
<td>Increased</td>
<td>Low</td>
<td>“Poor” item</td>
</tr>
<tr>
<td>8. slept badly</td>
<td>Sleep</td>
<td>0.524</td>
<td>4</td>
<td>3.209</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. tired</td>
<td>Misfit</td>
<td>0.003</td>
<td>5</td>
<td>21.708</td>
<td>Dropped</td>
<td>Low</td>
<td>“Poor” item</td>
</tr>
<tr>
<td>10. busy to take mind off things</td>
<td>Misfit</td>
<td>&lt; 0.0005</td>
<td>8</td>
<td>54.375</td>
<td>Dropped</td>
<td>High</td>
<td>Single item</td>
</tr>
<tr>
<td>11. hard to concentrate</td>
<td>Behavioural</td>
<td>0.189</td>
<td>7</td>
<td>9.989</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12. time passed slowly</td>
<td>Dejection</td>
<td>0.852</td>
<td>8</td>
<td>4.051</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13. change in appetite</td>
<td>Behavioural</td>
<td>0.426</td>
<td>7</td>
<td>7.024</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14. sad</td>
<td>Dejection*</td>
<td>0.016</td>
<td>8</td>
<td>18.855</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15. upset</td>
<td>Anxiety</td>
<td>0.282</td>
<td>9</td>
<td>10.907</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16. examined my breasts</td>
<td>Breast examination</td>
<td>0.152</td>
<td>4</td>
<td>6.711</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17. restless</td>
<td>Anxiety</td>
<td>0.262</td>
<td>9</td>
<td>10.039</td>
<td></td>
<td></td>
<td>Single item?</td>
</tr>
<tr>
<td>18. nervous</td>
<td>Anxiety</td>
<td>0.244</td>
<td>9</td>
<td>11.486</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Item Description</td>
<td>Dimension</td>
<td>Beta</td>
<td>Magnitude</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>-------------------------------------------------------</td>
<td>----------------------</td>
<td>------</td>
<td>-----------</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>19.</td>
<td>uneasy</td>
<td>Dejection</td>
<td>0.002#</td>
<td>8</td>
<td>24.413</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20.</td>
<td>taken long time to fall asleep</td>
<td>Sleep</td>
<td>0.383</td>
<td>5</td>
<td>5.277</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21.</td>
<td>examined my breasts in the mirror</td>
<td>Breast examination</td>
<td>0.012</td>
<td>4</td>
<td>12.891</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22.</td>
<td>withdrawn into myself</td>
<td>Behavioural</td>
<td>0.060</td>
<td>7</td>
<td>13.532</td>
<td></td>
<td></td>
</tr>
<tr>
<td>23.</td>
<td>unable to cope</td>
<td>Dejection</td>
<td>0.388</td>
<td>8</td>
<td>8.483</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24.</td>
<td>depressed</td>
<td>Dejection</td>
<td>0.228</td>
<td>8</td>
<td>10.564</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25.</td>
<td>symptoms from the breast (pins and needles)</td>
<td>Misfit</td>
<td>&lt; 0.0005</td>
<td>5</td>
<td>26.942</td>
<td>Increased</td>
<td>Low</td>
</tr>
<tr>
<td>26.</td>
<td>difficulty dealing work or other commitments</td>
<td>Behavioural</td>
<td>0.006#</td>
<td>7</td>
<td>19.862</td>
<td></td>
<td></td>
</tr>
<tr>
<td>27.</td>
<td>headache</td>
<td>Misfit</td>
<td>0.004</td>
<td>6</td>
<td>19.379</td>
<td>Increased</td>
<td>Low</td>
</tr>
<tr>
<td>28.</td>
<td>difficulty doing things around the house</td>
<td>Behavioural</td>
<td>0.310</td>
<td>7</td>
<td>8.261</td>
<td></td>
<td></td>
</tr>
<tr>
<td>29.</td>
<td>terrified</td>
<td>Anxiety</td>
<td>0.609</td>
<td>9</td>
<td>6.344</td>
<td></td>
<td></td>
</tr>
<tr>
<td>30.</td>
<td>taking sleeping tablets</td>
<td>Misfit</td>
<td>0.003</td>
<td>5</td>
<td>17.688</td>
<td>Increased</td>
<td>Low</td>
</tr>
<tr>
<td>31.</td>
<td>less interest in sex</td>
<td>Sexuality</td>
<td>0.828</td>
<td>5</td>
<td>2.154</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32.</td>
<td>not felt like having my breast caressed</td>
<td>Sexuality</td>
<td>0.428</td>
<td>5</td>
<td>4.905</td>
<td></td>
<td></td>
</tr>
<tr>
<td>33.</td>
<td>sick leave</td>
<td>Not included</td>
<td>Not included</td>
<td>Not included</td>
<td>Not included</td>
<td>Not included</td>
<td>High</td>
</tr>
</tbody>
</table>

Please notice: the item (no 33) asking about sick leave was not included in the Rasch analyses. *Item was deleted from the dejection dimension because of differential item function. #Marginal misfit also after a correction of Benjamini-Hochberg procedure."
Discussion

Six Rasch-fitting dimensions were identified encompassing 24 items. The dimensions cover: anxiety, behavioural impact, sense of dejection, impact on sleep, breast examination and sexuality. The six dimensions had sufficient sensitive to distinguish between groups *a priori* hypothesised to be different. The correlations between the six Rasch-fitting dimensions and the NHP were also as could be expected from a content point of view.

Two items not fitting the Rasch models were retained in the questionnaire because of high face validity. These were concerned with “feeling less attractive” and “kept busy to take mind off things”. An item “felt sad” belonged to the dejection dimension but had uniform DIF relative to diagnostic subgroups. The six dimensions and these three single items converged as expected with the emotional section of the NHP and diverged as expected with the pain and the physical mobility sections.

The sick-leave item was not included in the original Australian version of the PCQ. However, it seems to add useful information when measuring side effects and socioeconomic consequences of breast cancer screening.

Five “poor items” were identified. Four of these misfitted the Rasch model and had low face validity. Cronbach’s alpha also increased when they were deleted. The fifth “poor” item “being tired” showed an unclear picture. Although, it did not fit the Rasch model and had low face validity, Cronbach’s alpha dropped when it was deleted. Perhaps “being tired” describes a too general condition. Therefore, it is suggested that the five “poor” items should be deleted from the final instrument. It is worth mentioning that one of the “poor” items “keeping things from those who are close to you” is an item belonging to the original Australian version of the PCQ.

Collecting questionnaire data under two different conditions may result in biases. At Time I, the women completed the PCQ-DK33 at the recall clinic. Two weeks later at Time II the same women completed the PCQ-DK33 at home. Some women had the additional examinations at the recall clinic only one day after receiving the letter about the abnormal screening mammography. This short time interval made it necessary to complete the questionnaires at the recall clinic instead of receiving it by post. Women completing a questionnaire at a clinic may “smarten up” their answers to be polite. Therefore, the negative psychosocial consequences of an abnormal screening mammography would most likely be underestimated.

Three items showed differential item functioning relative to diagnostic subgroups. If an item functions differently in subpopulations other psychometric properties should determine the
“destiny” of the item. The qualitative study preceding the present study showed that the two items forming the breast examination dimension cover an important area in the context of abnormal and false-positive screening mammography. The results from the concurrent validity tests confirmed that the breast examination dimension measured something different from the other five dimensions. Consequently, it would be unwise to remove these items. However, special precautions should be taken when calculating scores of this dimension. The content of the item “being sad” is close to the content of the other five items in the dejection dimension. Therefore, it seems reasonable to delete this item if future studies continuously show DIF.

Three items from three different dimensions had problems with the order of thresholds. As seen in figures 2, 3 and 7 the disorder was caused by minor problems. If future studies show the same pattern it has to be decided if the response categories “A bit” and “Quite a bit” should be merged either by rescoring the items or re-designing the layout.

Two items belonging to two different dimensions showed a marginal misfit of probability to the Rasch model (no’s 19 & 26, table 4). However, the overall fit of the dimensions were satisfactory. Future studies including the questionnaire will show if these findings are consistent.

It was surprising that the sexuality dimension only distinguished between women having a normal and women having an abnormal screening mammography. This may indicate that the negative impact on sexuality after having an abnormal screening result had not declined or vanished one or two weeks after women were “free from” cancer suspicion.

The convergence between the dimensions of anxiety, behaviour and dejection and the emotional section of the NHP indicates some overlap between these three dimensions. Only longitudinal studies will reveal if this overlap is caused by redundancy. However, removing any of the dimensions would decrease content coverage. The lack of convergence between the two dimensions “breast examination” and “sexuality” and the emotional section of the NHP contradicts redundancy among the six dimensions.

The establishment of a traditional test-retest reliability coefficient requires at least two to four weeks where the condition for the respondents is stable. The condition for the women in Group 1 changed dramatically from Time I to Time II. At time I all women had been told that their screening result was abnormal. At time II nearly all women knew their diagnosis: breast cancer or false-positive. A satisfactory reliability of the measure was assessed with Cronbach’s alpha and Person Separation Index (table 1).
As shown in the Rasch analyses; four of five NHP sleep-items fitted the model. Two of the four items are content-wise equivalent to the two Rasch-fitting items of the PCQ-DK33. Therefore, adapting the two non-equivalent sleep-items from the NHP would probably add nuances to the sleep dimension of the new instrument.

As in many other questionnaires, the response options of the PCQ-DK33 are ordered categories. Several models for ordinal categorical responses have been suggested. The model used in the present study is the partial credit model (PCM) in which the item parameters are sometimes described as threshold parameters.\textsuperscript{34} The thresholds in the PCM may differ between items. In contrast, the rating scale model assumes that thresholds are homogenous across items apart from an additive factor describing the item difficulty.\textsuperscript{35} Rating scale models were considered during the analysis but abandoned because of lack of fit between the model and the observed item responses. The present study has shown the advantages of combining analyses based on IRT and CTT when assessing the psychometric properties of a questionnaire including dimensionality and “good” and “poor” single items. After establishing unidimensionality with the Rasch model CTT analyses were subsequently conducted. This order of analyses had several advantages: Firstly, more than half of the Rasch-misfitting items were confirmed also to be “poor” by the analyses of Cronbach’s alpha. Secondly, the internal consistency expressed as a Cronbach’s alpha coefficient was calculated only on items included in the Rasch-fitting dimensions. Thirdly, the results of testing known group validity and concurrent validity were only established on “good” items.
Conclusion and perspectives
In conclusion, the reliability and the construct validity of a condition-specific measure with high content validity for women having an abnormal screening mammography have been preliminary demonstrated. This new questionnaire covers the impact of experiencing an abnormal screening mammography on: anxiety, behaviour, dejection, sexuality, sleep and breast examination. In addition, the measure includes three single items covering: sick leave, feeling less attractive and kept busy to take mind off things.
The new instrument is currently in use in a major Danish survey and has been translated into Dutch, English and Norwegian. Future analyses on data from surveys will hopefully give an answer to the questions left from the present study:
- should all the three single items be retained in the final version of the measure?
- do the four items covering sleep form one dimension?
- will the item “felt sad” still have uniform DIF?
- will some items still have threshold that are not in order?

Acknowledgement
We are in great debt to the late professor Jill Cockburn for her inspiration and strong supported of developing a condition-specific measure of psychosocial consequences of abnormal and false-positive screening mammography. We want to thank the staff at the breast cancer screening clinic and the recall clinic at the Copenhagen University Hospital and at the Odense University Hospital for recruiting women to this study.
Reference List


10. **Discussion of article 3**

10.1. **Major findings**

Six Rasch-fitting dimensions encompassing 24 items covering; anxiety (6 items), behavioural impact (7 items), sense of dejection (5 items), impact on sleep (2 items), breast examination (2 items) and sexuality (2 items) were identified among the items of the PCQ-DK33 (appendix IV, page 136 & appendix V, page 141).

In addition two items not fitting the Rasch model were retained as single items in the questionnaire because of high face validity. One item “felt sad” belonged to the dejection dimension but had uniform differential item functioning relative to two of four diagnostic subgroups. Altogether these 27 “good” items plus a sick-leave item, measure different attributes of the same overall latent construct – the psychosocial consequences of abnormal screening mammography.

Five “poor items” were identified and should be deleted from the final instrument. For details see Table 4 in article 3 and the flow chart on page 103.

10.2. **Assessment of methods**

10.2.1. **Data collection**

At the screening recall clinics the women had a table to sit at when completing the questionnaires. In Copenhagen the table was placed in a corridor remote from the normal waiting area. In Odense, the women were offered a more private separate room. Despite the two different conditions there was no statistically significant difference in the degree of negative psychosocial consequences of abnormal screening mammography measured at the two recall clinics.

The difference in response rates between Group 1 and Group 2 may have different reasons:

- the motivation for completing the PCQ-DK33 might have been greater among women having an abnormal screening mammography compared to women having a normal mammography
- it might be easier not to participate in a survey when you are posted a questionnaire than when you are asked at a clinic
- the women in Group 2 were not posted a reminder

It would have been preferable to use the same administration procedures in the two groups and at the two different time points for Group 1.

The procedures had to be accepted due to logistical, practical, ethical and legal issues. Nevertheless, they might have caused biases. However, such biases would most likely minimise the differences actually found between the psychosocial consequences of abnormal screening mammography compared to consequences of normal screening mammography.

The women having a normal screening result were recruited according to which day and at which screening clinic they had an abnormal screening mammography. Women were invited to attend screening from the city of Copenhagen and from the council of Funen (Odense University Hospital) according to their permanent home addresses. At both screening centres the invitations are posted in clusters so that in the same period of time (days or weeks) the women aged 50 – 69 living in the same local area are invited. In the Funen council screening takes place in two different settings: at a screening centre at the University Hospital in the city of Odense and in a specially equipped screening-bus driving from one town to another and on the smaller Islands belonging to the council of Funen. There is no reason to believe that there would be any differences in income or education
levels between the groups and subgroups of women. The cluster-invitation procedure even makes differences in social groups less possible.

10.2.2. Statistical methods in general
Data were input manually by one person and proof-read by two other persons. The manual search for errors in the input of data showed 0.20% mistakes which were corrected before analyses. The main statistical method used to explore the psychometric properties of the PCQ-DK33 was the Rasch model (page 23). The Rasch model is a valuable tool for identifying dimensions and “poor” items of a questionnaire, for example when dealing with the items encompassing “sleep” in the present study:

10.2.2.1. Sleep items
Three items about sleeping problems were generated during the focus group interviews when the PCQ-DK33 was developed. While two of these items fitted the Rasch model the third item, “taking sleeping tablets”, did not. In the PCQ-DK33 the sleeping items have four response categories. In the NHP the items are dichotomous. The five sleeping items in the NHP were also analysed using the Rasch model. Four of the five NHP-sleeping items fitted the model. The content of the NHP-sleep item that did not fit the Rasch model was “taking sleeping tablets”. Therefore, the item “taking sleeping tablets” was regarded as a “poor” item as it did not measure the same construct as the other sleep items in both questionnaires. However, if the content of the item “taking sleeping tablets” had high face validity for women having an abnormal screening mammography the item could have been kept as a single item.

10.2.2.2. Response categories and threshold order
The Rasch model also explores the order of the response categories of the items. In the present study three items (no 28, 29 and 31) in three different dimensions (anxiety, behavioural impact and sexuality) had disordered response categories. The overall fit statistics of the three dimensions were sufficient. Therefore, the disorder of the response categories does not seem to be a matter of an illogical response pattern of the items. Item 28 and 29 are “extreme” items measuring the severest negative impact on the behavioural and anxiety latent variables respectively. It is possible that it is too difficult to differentiate between three nuances of impact when responding to extreme items; for example, distinguishing between “a bit”, “quite a bit” and “a lot” when asked if you have “felt terrified”. If only three items of 30 items have response categories that are not in order re-scoring these items is a better alternative than changing the layout of the questionnaire.

10.2.2.3. Differential item functioning
The reason for excluding 17 undiagnosed women from the DIF analyses was to avoid bias. The small number of women in this subgroup compared to the number of women in the other groups and subgroups could cause random errors. Furthermore, prior to the study no hypothesis was raised including this subgroup. Therefore, the group was also omitted from the test of known group validity. In previous surveys the impact on women of early recall has only been measured after final false-positive diagnosis and not while the women were still undiagnosed. If future studies include women on early recall and measurements are conducted in the critical period then DIF analyses including this subgroup will be possible.
Calculations of sum scores of a dimension can be adjusted if an item in the dimension has DIF. The adjustment includes four steps:
1) an item split is conducted
2) the mean score of the person location on this new virtual latent trait for the particular subgroup is assessed
3) the mean score of the person location on the new virtual latent trait for the remaining subgroups is assessed.
4) the locations are input in the “Equating Tests” feature of the RUMM2020 calculating new mean scores of each of the locations.
Such an adjustment will avoid over or underestimation of differences between groups.

10.2.2.4. Combining Item Response Theory and Classical Test Theory
The study has shown that there are several advantages to combining analyses based on Item Response Theory (IRT) and Classical Test Theory (CTT) when assessing the psychometric properties of a questionnaire.
After conducting the Rasch analyses and identifying the six dimensions and the seven items not fitting the Rasch model further analyses based on CCT were conducted. Hence, the CTT analyses were conducted based on the results of the Rasch analyses, where the Rasch model was considered a valuable “gold standard”.
Besides making the results of the CTT analyses more valid this approach also saves time. For example, assessing Cronbach’s alpha could be used to identify “poor” items. If the value of alpha increases or stays stable when an item is deleted from a pool of items then this particular item could be “poor”.
The seven items not fitting the Rasch model were re-checked for face validity by re-auditing the tape recordings from the focus group interviews. They were then deleted one at a time from the remaining items and Cronbach’s alpha was assessed before and after each deletion. The alternative would have been to delete all the 33 items one at a time and reassess Cronbach’s alpha for the 32 remaining items after each deletion.
The Cronbach’s alpha analyses confirmed the results from the re-audition of the tape recordings in six of the seven items. Only one item did not ‘behave’ as presupposed: “being tired”.
Two of the remaining seven non-fitting items were found to have especially high face validity for women having surgery and those on early recall. These were the items “less attractive” and “busy to take mind off things”. The content and meaning of these two items are described in article 2 (page 36) and discussed in details on page 50. Some of the 17 undiagnosed women at Time II were on a waiting list for surgical examination or early recall. The decision to retain or delete these two items for the version of the questionnaire was postponed until analyses based on larger subgroups of women having surgery and being on early recall could be conducted.

10.2.2.5. Test of concurrent validity
In the test of concurrent validity the sleep dimension of the new measure is shown to specifically measure an impact on sleep. The correlation with the sleep section of the NHP is 0.69, while the correlations with the remaining five dimensions are less than 0.5. The two dimensions “breast examination” and “sexuality” contribute to the psychosocial consequences of abnormal screening mammography differently from the other four dimensions. Their correlations with the sections of the NHP are different from the remaining four dimensions and are also different form each other respectively.
The correlation between the sections of the NHP and the three dimensions, “anxiety”, “behavioural” and “sense of dejection” are all similar. This might indicate redundancy. If longitudinal studies confirm this redundancy it would be possible to delete two of the three dimensions without losing the sensitivity of the instrument. However, women may feel that they are not being taken seriously if they are not asked about, for example, anxiety as this was an important topic for the women in the focus groups described in article 2 (page 36). Low response rate could in fact be a consequence of such a lack of content validity.

The three dimensions; “anxiety”, “behavioural” and “sense of dejection” encompasses altogether 18 items (or 19 items if “felt sad” is included in the dejection dimension) and the final instrument measuring psychosocial consequences of abnormal screening mammography will have a maximum of 30 items (the 27 “good” items, the sick-leave item and the two adapted NHP-sleep items).

Completing a questionnaire with 30 relevant items will not overload respondents and so there is no need for item reduction.

10.3. Justification of conclusion

Rosenbaum reports that items of a measure fitting the Rasch model possess criterion-related construct validity.91 Messick describes construct validity as the overarching concept of validity and that “almost any kind of information about a test can contribute to an understanding of its construct validity”.112 Messick continues: “fundamentally, all validation is construct validation, in the sense that all validity evidence contributes to (or undercuts) the empirical grounding or trustworthiness of the score interpretation”.112 For example both content validity and criterion validitya are pieces of evidence for construct validity.112 “Construct validation is an ongoing process”61 and thereby dynamic and can probably never be fully achieved. Putting all the pieces of validity-evidence together can form a picture of the construct validity of a measure.

It is suggested that the six Rasch-fitting dimensions encompassing 24 items, plus the three single items, plus the sick-leave item be retained in the final version of the questionnaire because together they have high content relevance and high content coverage in the setting of abnormal screening mammography. When testing these 27 items (the sick-leave item excluded) and the dimensions for concurrent validity, they diverge and converge as presupposed. The responsiveness of the dimensions and the single items and their ability to discriminate between known groups were found to be satisfactory.

In summary, the statement in the conclusion of article 3: “preliminary evidence of construct validity of a condition-specific measure for women having an abnormal screening mammography has been established” can be justified.

10.4. Contribution to the current knowledge

As mentioned on page 19 the question: “Assessing psychosocial/quality of life outcomes in screening: how do we do it better?” has recently been raised.76 The psychometric properties and quality of the new condition-specific instrument goes beyond the criteria “reliable measurement tools with high criterion and content validity” raised by the questioner cited above.76 Preliminary evidence of construct validity of the new instrument has been established. Therefore, the development and validation of a condition-specific instrument measuring psychosocial consequences of abnormal screening mammography could be regarded as a contribution to the research field of better assessing psychosocial consequences of screening.

---

a Concurrent validity and predictive validity are unified into the term criterion validity.61;76;112
The current study also contributes to the previous discussion on one or more domains of psychosocial consequences of false positives. The original Australian version of the negative part of the Psychological Consequences Questionnaire (PCQ) encompasses five items covering an emotional domain, four covering a social domain and three a physical.\textsuperscript{52} When developing and validating the PCQ Jill Cockburn suggested summing the raw scores for each domain, with a higher score indicating a greater dysfunction. Organisation of the items into three domains was qualitatively based i.e. in agreement with expert judges as to whether the postulated domains could be distinguished on the basis of item content.\textsuperscript{52;104} In three studies\textsuperscript{113-115} this qualitatively based scaling of the PCQ was not confirmed by exploratory factor analyses: a British study identified a single factor with an eigenvalue of 7.7 explaining 65\% of the variance,\textsuperscript{113} a Swedish study did not support the three underlying factors,\textsuperscript{114} and a Canadian study found two components\textsuperscript{115} with eigenvalues of 7.71 and 1.15.\textsuperscript{116}

The current study shows that having an abnormal screening mammography may have an impact on different psychosocial domains. Although the PCQ has a lack of content validity in a setting of abnormal screening mammography\textsuperscript{109} Cockburn was right in her qualitative approach that having an abnormal screening mammography can have an effect on different psychosocial domains. The reason why three studies cannot confirm these three domains in exploratory factor analyses could be that the negative part of the PCQ encompasses six ambiguous items.\textsuperscript{109}

The most frequently used questionnaires in a setting of false-positive screening mammography are: the General Health Questionnaire (GHQ), the Hospital Anxiety and Depression Scale (HADS), the PCQ and the State-Trait Anxiety Inventory (STAI).\textsuperscript{104} Rasch analyses have been conducted on data collected with the GHQ, HADS and the STAI.\textsuperscript{117-121} These studies found a difference between the “original” psychometric properties of the instruments when they were developed and the psychometric properties in new setting. The HADS has also been shown to have problems when analysed with confirmatory factor analyses in a new setting.\textsuperscript{122} In questionnaire surveys on consequences of false-positive screening mammography no Rasch analyses have been reported.\textsuperscript{104} The present study is the first study in a setting of screening mammography where analyses based on the Rasch model have been conducted.

Analyses of differential item functioning have regularly been used to check if item functioning is different after a questionnaire has been translated into another language.\textsuperscript{80} The most usual method to compared two groups in a questionnaire study is to compare the mean scores. It is important to eliminate items with DIF or alternatively adjust the mean score according to the DIF. Otherwise there is a possibility that the estimated differences between the groups are either too small or too large. In the present study DIF analyses was conducted relative to:

- the diagnostic subgroups
- the screening centres
- age of the women

to avoid bias when comparing mean scores.
11. Article 4
Development and validity of a condition-specific instrument measuring long-term psychosocial consequences of false-positive screening mammography

John Brodersen (MD, GP, PhD student) and Hanne Thorsen (MD, PhD, senior researcher) both Department and Research Unit of General Practice, University of Copenhagen, Denmark

Abstract
Objectives
The aim of this study is to develop a questionnaire with high content validity for long-term consequences of false-positive screening mammography and to test the construct validity of this new instrument.

Method
Content validity of the draft version of the questionnaire was assessed via six focus group interviews conducted with a total of 34 women. Questionnaire data were then subjected to analyses using Item Response Theory (IRT) and Classical Test Theory (CTT). Reliability, internal consistency, known group validity, construct validity and objectivity were established.

Results
Thirteen items with seven response categories each were generated during the group interviews covering long-term consequences of false-positive screening mammography. In the interviews some women remarked that it was difficult to choose between the response categories: “A little less” and “Less”, and between “A little more” and “More”. Rasch analyses confirmed that the same response categories had problems. The estimates of reliability were all acceptable.

The analyses using IRT and CTT revealed three strongly correlated dimensions. No single items or “poor items” were identified.

Conclusion
Preliminary evidence of the construct validity of a condition-specific measure with high content validity for long-term consequences for women having a false-positive screening mammography has been established. The new instrument covers the impact on the women’s existential values; relationship with their social network; being less or more relaxed/calm; and less or more anxious about breast cancer/belief in having or not-having breast cancer. A future reduction from seven to five response categories resulting in a new layout of the questionnaire is suggested.
Introduction

Breast cancer is a substantial health problem in the western world for example will every tenth woman in Denmark be diagnosed with breast cancer.\(^1\) Many women in western countries are therefore offered breast screening programmes to reduce mortality of the breast cancer.\(^2\) In EU biennial breast screening is recommended for women aged 50 – 69.\(^3\) The incidence of breast cancer diagnosed in each screening round is approximately 0.5\%.\(^4\) The specificity of biennial screening mammography has been reported as 97\%, \(^5\) with a positive predictive values of 13\%.\(^6\) A low positive predictive value results in a high rate of false positive screening results e.g. if a woman follows the EU recommendations, every fourth to fifth woman will have a false-positive screening mammography during 20 years of screening.\(^7\) In the UK, where three-yearly screening is offered to women aged 50 - 65, more than 50,000 women per year will have a false-positive screening mammography.\(^8\)

Several studies have investigated if women having a false-positive screening mammography have more negative short-term consequences than women having a normal screening mammography.\(^9\) Regardless of outcome measures, nearly all studies conclude that there are significant short-term adverse psychosocial effects of having an abnormal screening mammography, later confirmed as a false-positive screening mammography (hereafter referred to as a false positive).\(^10\) In contrast, there are few studies on long-term psychosocial consequences of breast cancer screening: in these studies the follow-up period varies from 1 - 35 months and the results are ambiguous.\(^9\);\(^10\)

Apart from long-term psychosocial consequences a false positive seems to have an impact on participation in subsequent screening rounds. In three studies the participation in the subsequent screening round was higher among women who had a false positive compared to those who had a normal result.\(^11\);\(^12\);\(^13\);\(^14\) In another study it was lower \(^15\) and in one study the frequency of participation was unaffected.\(^16\) The use of health-care services has also been shown to increase among women after having had a false positive.\(^17\) Finally, an increased frequency of breast self-examination has been reported as a long-term consequence of false positives.\(^18\);\(^19\);\(^20\);\(^21\)

Questionnaire surveys have measured different aspects of long-term psychosocial consequences after a false positive.\(^14\);\(^20\);\(^22\);\(^23\);\(^24\);\(^25\);\(^26\);\(^27\) Some of these studies found that women having a false positive had measurable adverse psychosocial consequences at 5 - 35 months.\(^14\);\(^21\);\(^24\);\(^26\);\(^27\) Other studies found no long-term consequences\(^22\);\(^25\);\(^28\) and the interpretations of results in two studies are unclear.\(^20\);\(^23\)

McCaffery and Barrat underline the importance of high content validity when using questionnaires to measure the psychosocial outcome of screening.\(^29\) None of the articles on long-term
consequences of false positives report any pre-test of content relevance and content coverage of the questionnaires. It is therefore likely that the questionnaires were inadequate in measuring validly the long-term consequences of false positives and this may explain the ambiguity of the results in the surveys.

A study on content validity of the Danish version of the Psychological Consequences Questionnaire (PCQ) in measuring the psychosocial consequences of abnormal screening mammography has recently been conducted (article 2 of this thesis). In this study it was revealed that women’s experiences differ from one period to another. Women stated that their experiences in the critical period from abnormal screening mammography until final false-positive diagnosis differ entirely from those of the period after the final diagnosis. The women argued that some issues could be raised only after they had been declared “free from” cancer suspicion. Issues exclusively relevant to the period after final diagnosis were classified as long-term psychosocial consequences of false positives.

The first aim of the present study is to investigate these consequences and develop a condition-specific questionnaire covering long-term psychosocial consequences of false-positive.

When summing raw scores of items in a scale, an assumption of unidimensionality is made i.e. that the items describe different aspects of the same construct and can be added. When the response options are categorical on an ordinal scale as in many questionnaires Item Response Theory (IRT) models can be used to assess the psychometric properties of the questionnaire. The Rasch models (an IRT model) are a formal representation of perfect measurement. Where items are shown to fit a Rasch model the measure can be shown to posses criterion-related construct validity, to be objective, sufficient and, therefore, also reliable. For these reasons the Rasch model is considered a valuable “gold standard” against which measures should be compared. Item analyses by Rasch models explore in depth the degree to which the requirements of construct validity are met. Items are assumed to monotonically relate to one dimension, they are assumed to be locally independent and it is assumed that there is no differential item functioning. The sufficiency of the model support computation of the raw scores. IRT analyses also explore how the items included in each dimension are interrelated and ordered on a latent trait.

The second aim of this study is to explore the psychometric properties of a new condition-specific instrument covering long-term psychosocial consequences of false-positive using IRT and CTT analyses.
Material and methods

Development of items

Six focus group interviews were conducted with women who had been screened for breast cancer in the year prior to the interviews. The women had experienced that their screening mammography was abnormal and later false-positive. Details are reported in article 2 on how the participants were grouped in the six interviews and how the interviews were conducted.

Data collection

Data were collected at the screening centre of the Copenhagen University Hospital from 24.01.2003 to 27.02.2003.

After conducting the six focus group interviews a draft version of the questionnaire was posted to women who had had a false positive in the previous 1 – 6 months (Group A - Time I). Women were asked to complete the test questionnaire and return it in an enclosed stamped addressed envelope. If they had not responded after two weeks, they were posted a reminder including a covering letter, a test questionnaire and a stamped addressed envelope.

Two weeks after having completed the test questionnaire at Time I, the same women were again posted the same questionnaire (Group A - Time II), asked to complete it and return it in an enclosed stamped addressed envelope. This second data set with the data set from Time I was used to assess the test-retest reliability.

For each woman who had received the test questionnaire at Time I (Group A), another two women were posted the test questionnaire and asked to return it completed in an enclosed stamped addressed envelope (Group B). These women had had a normal screening mammography at the same time as the women in Group A, Time I. It was hypothesised that women with a false-positive screening result would experience more long-term psychosocial consequences compared to those with a normal screening result (known group validity).\textsuperscript{37,38} For ethical and legal reasons the draft version of the questionnaire was posted to the women in Group B by the screening clinics anonymously and only the age of the participants was disclosed for the researchers. Therefore it was not possible to send reminders to the women in Group B.

In Group A, 127 of 150 women (84.7%) returned the test questionnaire at Time I. At Time II, 105 (82.7%) of the women in Group A returned the test questionnaire. In Group B, 300 women received the questionnaire by post and 197 (65.7%) returned the questionnaire. There were no statistically significant differences in mean and range of age between the women in the three groups.

Statistical analyses
Item responses were analysed by the conditional distribution of items given total person scores in order to avoid assumptions on the distribution of the latent trait being measured. The pairwise estimation procedure implemented in software program RUMM2020 was used to estimate the item parameters.\textsuperscript{39,40} The analysis of the fit of item responses to the Rasch model were based on analyses of residuals comparing observed to expected item responses both for separate individuals and for different score groups. The overall fit of the model was assessed by the Wright-Panchapakesan $\chi^2$ statistic summarising standardised residuals over score groups and items.\textsuperscript{41} Item fit statistics summarising standardised residuals in different score groups were used to identify misfitting items. Data from Group A, Time I and Time II and data from Group B were pooled for IRT analyses. Differential item functioning (DIF) relative to person covariates was checked by analyses of variance examining the degree to which individual residuals for specific items depended on the covariates. Absence of evidence of interaction between the covariates and the estimated trait parameters were taken as evidence that DIF was uniform. Finally, the assumption of local independence was checked by examination of the degree to which individual residuals were correlated.

Reliability was assessed by Cronbach’s alpha defining the lower bound for the test-retest correlation of the raw scores\textsuperscript{42,43} and by the so-called Person Separation Index calculating the lower bound for the test-retest correlation of the estimated values of the latent trait being measured.\textsuperscript{44} The test-retest reliability was assessed by the Pearson correlation between Time I and Time II for Group A. Data were analysed with CTT using the software SPSS for Windows version 13.0 and the software Mplus 2.14 for confirmatory factor analysis.\textsuperscript{45} For IRT analyses the software RUMM2020 was used.\textsuperscript{39}

The study was approved by the local ethical committee.

Results

Development of items

Inspired by the ten items in the PCQ covering positive aspects of participation in breast cancer screening the women in the first focus group discussed their experiences after final diagnosis. They agreed that the consequences after final diagnosis were not only positive. Compared to the time before participation in screening, the women argued that consequences after final diagnosis should be described as on-going changes e.g. negative as well as positive changes in their “relationship to friends”.
In the first focus group it was revealed that the subject matter of the question in seven of the ten positive PCQ-items was relevant as long-term psychosocial consequences of false positives. The subject matter of these seven relevant questions was:

- Improved relationship to friends or relations
- Getting on better with those around you
- Feeling more relaxed
- A greater sense of well being
- Hopeful about future
- Reassurance of not having breast cancer
- Less anxious about breast cancer

The question: “Improved relationship with friends or relations” was considered double-barrelled and was split into two questions. However, the subject matter of the resulting eight questions was insufficient to cover all long-term psychosocial consequences of false positives. The women in the first focus group suggested five more issues to be covered. These issues were confirmed in the following group interviews: accordingly, five new questions were generated (table 1).

<table>
<thead>
<tr>
<th>1 belief that I do not have breast cancer</th>
<th>8 thoughts about the future</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 anxiety about breast cancer</td>
<td>9 enjoyment of life*</td>
</tr>
<tr>
<td>3 feeling relaxed</td>
<td>10 value life*</td>
</tr>
<tr>
<td>4 feeling calm*</td>
<td>11 awareness of life*</td>
</tr>
<tr>
<td>5 relationship with my family</td>
<td>12 thought about the broader aspects of life*</td>
</tr>
<tr>
<td>6 relationship with friends</td>
<td>13 sense of well-being</td>
</tr>
<tr>
<td>7 relationship with other people</td>
<td></td>
</tr>
</tbody>
</table>

* new items generate from focus group interviews

The response options from the PCQ were found irrelevant and incomprehensive when asking about long-term psychosocial consequences. In the first focus group the women suggested three response categories: “No”, “The same as before” and “Yes”. This was confirmed in the second focus group: however, the participants in this group found “No” and “Yes” too categorical. Therefore the number...
of response options was extended from three to seven and presented to the participants in the subsequent focus groups (table 2).

Table 2. Seven response categories

<table>
<thead>
<tr>
<th>Much less</th>
<th>Less</th>
<th>A little less</th>
<th>The same as before</th>
<th>A little more</th>
<th>More</th>
<th>Much more</th>
</tr>
</thead>
</table>

The women who participated in the last four focus groups found that the seven response options were relevant and allowed differentiated answers.

As part of the focus group interviews the participants completed the test questionnaire. After this exercise, some women remarked that it was difficult to choose between “a little less” and “less”, and between “a little more” and “more”. Conversely, it seemed that selecting “much less”, “much more”, or “the same as before” was easier. Despite the contradictory statements, it was decided to statistically test the questionnaire with seven response categories following each of the thirteen items. In addition the thirteen items were re-worded to fit the seven response options.

*Rasch analyses on the data of the draft version of the new questionnaire*

The response pattern showed that between 0 – 11% responded to one of the three response categories “much less”, “less”, and “a little less” (table 2). It was therefore decide to perform Rasch analyses after transforming the data. The data from the three categories were recoded in three different ways:

- as missing data
- as if they all were a response to the category “the same as before”
- laterally reversed, therefore a response to “much less” became a response to “much more”, “less” became “more”, and “a little less” to “a little more”.

Rasch analyses were conducted independently on each of the three recoded data sets. Irrespective of how the data were transformed, the differences of the results of the Rasch analyses were as described below:

- The thirteen items did not form a unified dimension
- The response category “a little more” had problems in ten to twelve of the thirteen items
- In item 6 (relationship with friends) the two response categories “a little more” and “more” had problems in two of the three transformed datasets
- Item 12 was the only item where all the four response categories worked sufficiently in all the analyses.

In a Category Probability Curve the likelihood of a response can be compared with the person location. If a response category is informative and works well, an “open window” should appear in the Category Probability Curve. In this “open window” the likelihood of responding to the category is greater than responding to the remaining categories. As an example, figure 1 shows how the two response categories in item 6 “a little more” and “more” did not work. Figure 2 shows how these two response categories were able to work sufficiently and give additional information compared to the other two response categories after they were collapsed.

Figure 1. Category Probability Curve of item 6 ("relationship with friends” see table 3) before collapsing response categories

![Category Probability Curve](image)

The logit scale from -2 to +4 on the x-axis symbolises the latent trait of long-term psychosocial consequences of false-positives, with the severity increasing towards the right. 0 “the same as before”, 1 “a little more”, 2 “more” and 3 “much more".
The logit scale from -2 to +4 on the x-axis symbolises the latent trait of long-term psychosocial consequences of false-positives, with the severity increasing towards the right. 0 “the same as before”, 1 “a little more” and “more” and 2 “much more”.

As a consequence of the Rasch analyses it was decided to collapse the two response categories “a little more” and “more” in all thirteen items and rescore accordingly. The results of the Rasch analyses on the three rescored data sets were:

- The response categories in all thirteen items were in order
- The thirteen items still failed to form a unified dimension
- Six items covering an impact on existential values formed one dimension (item 8 - 13, table 2)
- Three items covering an impact on social relations formed a second dimension (item 5 – 7, table 2)
- Two items about being relaxed or calm formed a third dimension (item 3 + 4, table 2)
- Two items describing anxiety and reassurance about breast cancer formed a fourth dimension (item 1 + 2, table 2)

The Wright-Panchapakesan Chi Squared fit Statistics, the Person Separation Index and the Cronbach’s alpha of the four dimensions fitting the Rasch model are listed in Table 3. This information is given with four response categories and with three response category (before and
after collapsing the two response categories “a little more” and “more”) and for all three recoded data sets.

TABLE 3 ABOUT HERE

Item no 13 “sense of well-being” had uniform DIF relative to diagnostic subgroup in all three data sets and so had item no 8 “thoughts about the future” in one data set.

To estimate how the four dimensions were correlated a subsequent confirmatory factor analysis was conducted. The analysis revealed a latent trait behind the three dimensions: “social relations”, “being relaxed or calm” and “anxiety and reassurance about breast cancer” with p-values in the three recoded and rescored data sets between 0.12 – 0.42.

Know group validity
In all four dimensions there was a statistically significant difference, insofar as women with a false positive had experienced more long-term psychosocial consequences than women with a normal screening result. P-values were less than 0.0005. The distribution of the relative frequencies in the five response categories (with the response categories “a little less” and “less” collapsed and “a little more” and “more” collapsed) in the existential dimension are shown in figure 3.

Figure 3. The relative frequencies in the existential dimension

```
```

“Normal”, the women that had a normal screening mammography (group B) and “False positive”, the women who had a false-positive screening mammography (Group A, Time I)
Table 3.

<table>
<thead>
<tr>
<th>Dimensions (No. of items)</th>
<th>WP $\chi^2$ Degrees of freedom</th>
<th>p</th>
<th>Person Separation Index</th>
<th>Cronbach’s alpha</th>
<th>WP $\chi^2$ Degrees of freedom</th>
<th>p</th>
<th>Person Separation Index</th>
<th>Cronbach’s alpha</th>
<th>WP $\chi^2$ Degrees of freedom</th>
<th>p</th>
<th>Person Separation Index</th>
<th>Cronbach’s alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data set with four response categories</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Breast cancer (2)</td>
<td>17.71</td>
<td>8</td>
<td>0.023</td>
<td>0.82</td>
<td>0.84</td>
<td>12.61</td>
<td>8</td>
<td>0.126</td>
<td>0.75</td>
<td>0.79</td>
<td>19.83</td>
<td>8</td>
</tr>
<tr>
<td>Relaxed/calm (2)</td>
<td>16.46</td>
<td>9</td>
<td>0.058</td>
<td>0.92</td>
<td>0.91</td>
<td>19.14</td>
<td>9</td>
<td>0.024</td>
<td>0.92</td>
<td>0.91</td>
<td>18.66</td>
<td>9</td>
</tr>
<tr>
<td>Social relations (3)</td>
<td>11.65</td>
<td>14</td>
<td>0.634</td>
<td>0.94</td>
<td>0.91</td>
<td>11.46</td>
<td>14</td>
<td>0.649</td>
<td>0.94</td>
<td>0.91</td>
<td>26.45</td>
<td>11</td>
</tr>
<tr>
<td>Existential values (6)</td>
<td>52.79</td>
<td>36</td>
<td>0.035</td>
<td>0.93</td>
<td>0.93</td>
<td>59.99</td>
<td>36</td>
<td>0.007</td>
<td>0.93</td>
<td>0.93</td>
<td>54.40</td>
<td>36</td>
</tr>
<tr>
<td>Data set with three response categories</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Breast cancer (2)</td>
<td>7.10</td>
<td>5</td>
<td>0.213</td>
<td>0.76</td>
<td>0.82</td>
<td>11.90</td>
<td>4</td>
<td>0.018</td>
<td>0.72</td>
<td>0.76</td>
<td>6.56</td>
<td>4</td>
</tr>
<tr>
<td>Relaxed/calm (2)</td>
<td>3.61</td>
<td>5</td>
<td>0.607</td>
<td>0.86</td>
<td>0.91</td>
<td>3.14</td>
<td>5</td>
<td>0.679</td>
<td>0.87</td>
<td>0.90</td>
<td>3.89</td>
<td>5</td>
</tr>
<tr>
<td>Social relations (3)</td>
<td>17.15</td>
<td>10</td>
<td>0.071</td>
<td>0.92</td>
<td>0.91</td>
<td>16.43</td>
<td>9</td>
<td>0.058</td>
<td>0.92</td>
<td>0.91</td>
<td>10.75</td>
<td>11</td>
</tr>
<tr>
<td>Existential values (6)</td>
<td>53.41</td>
<td>36</td>
<td>0.031</td>
<td>0.91</td>
<td>0.92</td>
<td>62.27</td>
<td>36</td>
<td>0.004</td>
<td>0.91</td>
<td>0.91</td>
<td>58.42</td>
<td>36</td>
</tr>
<tr>
<td>Existential values (7)*</td>
<td>58.26</td>
<td>42</td>
<td>0.049</td>
<td>0.92</td>
<td>0.92</td>
<td>68.01</td>
<td>42</td>
<td>0.007</td>
<td>0.91</td>
<td>0.91</td>
<td>62.41</td>
<td>42</td>
</tr>
<tr>
<td>Existential values (8)#</td>
<td>63.82</td>
<td>48</td>
<td>0.063</td>
<td>0.92</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*after splitting item 13 according to the uniform DIF found, #after splitting item 8 according to the uniform DIF found
Test-retest reliability

The test-retest correlation coefficient with seven response categories varied on the single item level from 0.64 – 0.79.

Discussion

In contrast to the study on the development of the original version of the PCQ, where women who had undergone surgery or early recall were not included, these groups of women were included in the present study. The women in the group interviews stated that some issues concerning long-term psychosocial consequences of false positives were not covered by the PCQ. However, the long-term consequences already covered by some PCQ-items had to be re-worded and the response options re-designed. The subject matter of the items was maintained but the final version of the test questionnaire was different from the PCQ and therefore was regarded as a new instrument.

The women in the focus groups emphasised that some questions could be asked only after final diagnosis. This is in agreement with the result obtained during the development of the PCQ and with a study on quality of life following a false positive. It is noteworthy that the subject matter of the five new items was generated during the first focus group interview with participants who had undergone the least distressing additional examinations. The fact that the subject matter of these items was confirmed in the following five group interviews and that no new issues were raised after the first focus group indicates a high degree of data saturation and thereby a high content validity of the new instrument.

In a questionnaire it is essential to avoid asking questions that are irrelevant. On the other hand, it is also important to include all relevant issues revealed in the qualitative pre-test of the questionnaire. It was therefore decided to include response options in each item indicating “no change” and changes in two opposite directions so that, for example, indicating either more or less confidence in not having breast cancer was made possible in the same item. This decision was taken to comply with the results obtained in the focus group interviews. The decision is supported by results of a study showing that women diagnosed with breast cancer experience both positive and negative consequences of their disease. Another reason for designing the new instrument with response options including changes in two opposite directions was to avoid complicated layout of the questionnaire e.g. skip patterns and mutually exclusive items.

However, this way of constructing a questionnaire demands a particular item scoring system. It is important not to score the changes so that they neutralise each other: therefore, an assessment of a
mean score of the dimensions do not express the actual distribution of changes. Instead, the range of the relative or absolute frequencies in all the response categories of the dimensions should be assessed.

It was planned to apply IRT analyses together with analyses of CTT to optimise the psychometric validation of the instrument. The Rasch model presupposes that “changes” occur in one direction only. Because only 0 – 11% of responses covered the three categories: “a little less”, “less” and “much less” it was decided as a compromise to transform the data so analyses based on the Rasch model were possible.

Although some of the women in the focus groups were reluctant to accept as many as seven response categories, it was decided to maintain these categories in the test version of the questionnaire. The Rasch analyses confirmed the findings from the interviews that the women had difficulties distinguishing between “a little less/more” and “less/more”. Therefore, both qualitative and quantitative results support a future reduction in the number of response options, resulting in a new layout of the questionnaire. Although the reliability estimates found in the present study are all acceptable, a future reduction in the number of response options may also enhance the reliability of the re-designed instrument.

Only weak evidence was disclosed against the six existential items (no. 8 – 13, table 1) forming one dimension (p-values from 0.004 – 0.06 in table 3). Five of the six existential items showed sufficient fit to the Rasch model in all the analyses. However, item no 10 “value life” showed a marginal misfit of probability to the Rasch model in all three recoded data sets (with four response categories) and also in the three rescored data sets (with three response categories). If item no 10 was removed from the existential dimension the overall fit of the dimension increased so that p > 0.05 in all the analyses. This might indicate that item no 10 has to be regarded as a single item not belonging to the existential dimension. Future studies with the previous suggested new layout of the questionnaire will show if these findings are consistent.

Studies to come with the re-designed questionnaire will also show if item no 13 “sense of well-being” consistently has DIF. If so, item no 13 could be deleted from the final version of the measure. Alternatively, the scores in the existential dimension have to be adjusted according to the apparent DIF so differences between diagnostic subgroups are not under or overestimated.
Conclusion
The construct validity of a condition-specific measure with high content validity for long-term psychosocial consequences for women having a false-positive screening mammography has preliminary been established. This new instrument encompasses four dimensions covering the impact on the women’s existential values; relationship with their social network; being less or more relaxed/calm; and less or more anxious about breast cancer/belief in having or not-having breast cancer. The last three dimensions are strongly correlated with a latent trait behind them.

Perspectives for future research
For measuring psychosocial consequences of screening in the setting of breast cancer a two-part valid condition-specific instrument, COS-BC (Consequences of Screening on Breast Cancer), measuring:
1. psychosocial consequences of abnormal screening mammography (part 1)
2. long-term psychosocial consequences of false positives (part 2)
is now available.
The COS-BC has been translated into Dutch, English and Norwegian and is currently in use in a major Danish survey. Data from this survey and from studies from other countries will allow statistical re-evaluation of the psychometric properties of the re-designed part 2. Hopefully these analyses will answer the questions remaining from the present study concerning part 2 of the COS-BC:
- are five, instead of seven, response categories sufficient?
- will the reliability and internal consistency of the re-designed questionnaire be improved?
- are the remaining psychometric properties of the part 2 of the COS-BC consistent with the results of the present study after the changes in layout?
- will item no 13 “sense of well-being” have DIF?
- will item no 10 “value life” belong to the existential dimension or will the item have to be regarded as a single item?

Another important target would be to establish the minimal important difference of the subscales in the dimensions of the COS-BC including the single items.
Some items and/or some of the dimensions of the COS-BC are specific for women participating in breast cancer screening. There is a need for valid instruments with high content validity and good psychometric properties in the research area of psychosocial consequences of cancer.
The relevance of the COS-BC, without the breast cancer specific items, is currently being explored in the setting of lung cancer screening. The aim is to establish core questions relevant for any kind of cancer screening. The core-questionnaire, COS, should then have specific items added, depending on which cancer the screening programmes are targeted at.

Acknowledgement
We would like to thank MSC, associate professor Svend Kreiner, Department of Biostatistics, Institute of Public Health, University of Copenhagen, for his great help with the interpretation of the statistical analyses. We also thank the staff at the breast cancer screening clinic and the recall clinic at the Copenhagen University Hospital for recruiting women to this study.
Reference List


51. Rimer BK, Bluman LG. The psychosocial consequences of mammography. [Review] [27 refs]. *Journal of the National Cancer Institute* 1997;Monographs.:131-8.
12. Discussion of article 4

12.1. Major findings
During focus group interviews thirteen items with seven response categories each were generated covering long-term psychosocial consequences of false-positive screening mammography (hereafter referred to as a false positive). The long-term psychosocial consequences differ entirely from those in the critical period from abnormal screening mammography until final false-positive diagnosis and could be asked about only after the women had been declared “free from” cancer suspicion. In the interviews some women stated that is was difficult to choose between the response options “A little less” and “Less” and between “A little more” and “More”. Rasch analyses confirmed that the same response categories had problems. Therefore, a future reduction from seven to five response categories resulting in a new layout of the questionnaire is suggested.

The new instrument encompassed four dimensions covering:
• the impact on the women’s existential values
• the relationship with their social network
• being less or more relaxed/calm
• less or more anxious about breast cancer/belief in having or not-having breast cancer

The last three dimensions were shown to be strongly correlated in a confirmatory factor analysis.

12.2. Assessment of methods
Translation and adaptation
The pitfalls of translating and adapting questionnaires from one language to another are discussed previously on page 49. For detailed information of the translation and adaptation process see appendix III (page 123).

12.2.1. Assessment of content validity
The details of how the focus group interviews were prepared and conducted are described and discussed in article 2 (page 36) and on page 50.
In the open-ended discussion of the first group interview it became obvious that a discussion of the psychosocial consequences of an abnormal and false-positive screening mammography had to be divided into two separate discussions. This was confirmed in the following five group interviews.

The discussion of the psychosocial consequences of the critical period from abnormal screening mammography until final false-positive diagnosis is reported in article 2 and in the following discussion of the paper.

The women in the first group also stated that the psychosocial consequences experienced after receiving a known false-positive diagnosis was of a more abstract nature compared to the experiences they had had in the critical period. When the women were given the translated Danish version of the PCQ they all agreed that the subject matter of some of the positive items of the PCQ were relevant from their own experiences after final false-positive diagnosis.

Two of the original positive PCQ-items (no’s 3 and 8, appendix III, page 123) describe the women’s relationship to their social network. Item 3 “Improved relationship with friends or relations” was found to be double-barrelled and was therefore divided into two items (appendix III, page 123). This resulted in three items covering social relationships. The women in the interviews wanted these three items on the same page of the questionnaire and in a particular order. They
recommended that the item about relationship to family had to precede the item about relationship with friends and the item about relationship to other people should come last. If the items about family and friends did not precede the item about other people then the women found the last item hard to interpret. The meaning of “other people” was unclear if this item was asked before the other two social network items. Presenting the three social items in this particular order resulted in an interpretation of “other people” as people that the women were not close to such as neighbours, colleagues, the hairdresser etc.

As the psychosocial consequences experienced after final false-positive diagnosis were of a more abstract nature than in the critical period, the overall introduction to the positive part of the PCQ: “All things considered, would you say your experiences at the Breast X-ray Program have caused any of the following” was adapted into Danish. It was not relevant to set a time frame of, for example, “over the last week” as used for the negative PCQ-item (appendix IV & V) and in the PCQ-DK33 (article 2 and 3, page 36 and 52).

Three positive PCQ-items were the same as three negative PCQ-items, but the meaning was reversed. The content of these items are:

1. Feeling more able to do things which I normally do
2. Feeling more able to meet my home and/or work responsibilities
3. Been sleeping better

These three original positive PCQ-items were excluded from the instrument because:

Re. 1) Firstly, the women in the group interviews stated that the three items were not relevant for the period after final diagnosis. Some of them stated that they slept better than in the critical period but none of the women stated that they slept better than before they were screened.

Re 2) Secondly, the three items could be misunderstood because the women did not know if they were supposed to compare the statements with the time before screening or with the time in the critical period.

Re 3) Finally, it is problematic to have two items in the same questionnaire asking about the same issue where one item is asking positively and the other negatively. It is most likely that some respondents will give opposing answers. Therefore, the validity of such items are questionable. Other respondents might be suspicious that a hidden agenda exists such that the real interest is not in the person’s answer but more a test of the person’s ability to answer consistently. This might result in a reduced response rate.

12.2.2. Data collection

The difference in response rates between Group A (82.7% - 84.7%) and Group B (65.7%) may be the result of the following reasons:

- the motivation for completing the questionnaire might have been greater among women having a false positive compared to women having a normal mammography
- the women in Group B were not posted a reminder

It would have been preferable to have posted a reminder to both groups. However, this was not feasible for practical, ethical and legal reasons. The different procedures in the two groups had to be accepted despite the fact that it might have caused biases. Future surveys where the instrument is used (and where it will be possible to send reminders to all the participants) will show if the differences in long-term psychosocial consequences actually found in the present study were biased.

A copy of the 13-item questionnaire with seven response categories posted to the women is found in appendix VI, page 146 and appendix VII, page 151.
12.2.3. Statistical methods in general

Data were input and proof-read by two persons. The manual search for errors in the input of data showed 0.14% mistakes that were corrected before analyses. The response pattern showed that between 0 – 11% responded to one of the three response categories “much less”, “less”, and “a little less”. However, 28 of the 127 (22%) women that returned the questionnaire in Group A, time I had affirmed at least one of the three response options in at least one item.

The data were recoded in three different ways to allow Rasch analyses to be conduct. Confirmatory factor analyses (CFA) were also conducted on the recoded raw scores which confirmed the existence of three of the four dimensions (with p-values greater than 0.05). The existential dimension was not confirmed by this process. After collapsing the two response categories “a little more” and “more” CFA was repeated. Again this provided evidence for three dimensions with p > 0.05 but the forth, the existential dimension, did not fit the model. This misfit of the existential items in the CFA might be caused by:

- the disorder of the response categories in the three recoded data sets
- the differential item functioning (DIF) found in item 13
- the misfit of item 10 in the existential dimension

When item 10 was deleted from the existential dimension and CFA repeated on the recoded and rescored data sets, the resulting p-value was between 0.0126 – 0.0501 with a model of a latent factorial structure supporting the five remaining existential items. This, together with the Rasch analyses, indicates that item 10 should either be regarded as a single item or that it should belong to the existential dimension. However, it is premature to conclude on this issue because the evidence is based on data that have both been recoded and rescored. Therefore, further decisions of the destiny of item 10 will be taken after a re-evaluation of the psychometric properties of the re-designed questionnaire with five response categories.

12.2.3.1. Order of response categories

In most of the items the response categories were not in order when Rasch analyses were conducted on data with four response categories (before recoding the data which corresponded to seven response categories). After collapsing the four response categories to three (which corresponded to a reduction from seven to five categories) the probabilistic fit to the Rasch model changes in the four dimensions respectively (Table 3 in article 4). In most of the cases the probability of fit increased but in some instances it decreased. Except for the existential dimension none of the p-values decreased to a level of misfit. This indicates that the reason for the disorder of the response categories is not a matter of an illogical response pattern between “a little less” and “less” and between “a little more” and “more”. A more plausible reason is that the differences in the quantity of information between “a little less” and “less” and between “a little more” and “more” is too small compared to the differences in the quantity of information between “the same as before” and the two extreme options “much less” and “much more”. This confirms the results from the group interviews where some of the women had difficulties choosing between the response options “a little less” and “less” and between “a little more” and “more”.

12.3. Justification of conclusion

The content of all the thirteen items in the instrument were generated during the first focus group and confirmed in the following five group interviews. As no new issues were raised after the first focus group, this indicates a high degree of data saturation and hence a high content validity of the new instrument.
Between ten and twelve of the thirteen items had disordered response categories. After collapsing the response categories from seven to five categories the disorder disappeared and the probability to the Rasch model increased in most of the analyses. These facts, together with the information from the qualitative interviews support the reduction in numbers of response options as suggested and the necessity of a new layout of the questionnaire.

The estimates of reliability were all acceptable, with relative high Cronbach’s alphas, high Person Separation Indexes and acceptable test-retest coefficients. A reduction in response categories will probably result in a higher reliability.

No local dependency among the items was identified in the Rasch-fitting dimensions.

Combining the validity-evidence provides preliminary justification for asserting that evidence of the construct validity of a condition-specific measure with high content validity for long-term consequences of false positives has been established.

12.4. Contribution to the current knowledge

In a study from 1990 Gram et al. reported long-term consequences of false-positive screening mammography such as: 1) pain and reduces sexual sensitivity of the breast after surgery, 2) anxiety about breast cancer and 3) finding ones life more precious. Two other studies from 1986 and 1990 reported higher levels of worry and anxiety about breast cancer as a long-term consequence of false positives and one woman stated she had reduced sexual sensitivity and pain of the breast after surgery. The present study confirms these findings.

It is a well known that after traumas and existential crisis people may change their view on values and meaning of life. These changes can be interpreted by the individual as positive and negative or a combination of both. This has been seen in a qualitative study on Danish cancer patients’ reactions one or more years after diagnosis. Antonovsky also reported similar complex reactions in interviews with former Jewish victims of the Second World War who had spent time in concentration camps. Antonovsky described a concept of sense of coherence as a salutogenic model on the relationship between health and disease. He found that the victims either had greater or lesser sense of coherence as a consequence of their traumas. Antonovsky did not use the positive reactions or greater sense of coherence as justification for sending people to concentration camps. In comparison, Gram et al. reported an overall positive impact on life as long-term consequences of false positives. However, the authors emphasised that it would be unreasonable to put these positive changes on the positive side of the balance sheet of breast cancer screening, “since first the fear, then the relief, are induced by the same screening”. In the present study women stated that the long-term impact on different aspect of life could be positive as well as negative. The instrument was designed so the response options were comprehensive to these statements and it was made possible to affirm an item both in a positive or negative direction. A change in either the positive or the negative direction from the response option “the same as before” must be seen as a psychosocial consequence of screening. The test of known group validity in the current study showed that there was a statistical significant difference between the group of women having a normal and a false-positive screening mammography. Women having a false positive had experienced greater changes and more impact on existential values, in the sense of being calm/relaxed, in relationship to their social network and in their anxiety/belief in having or not-having breast cancer. These changes, whether positive or negative, were the result of a process. Firstly, the abnormal screening mammography raised uncertainty and fear which was followed by the final false positive diagnosis. After being declared “free from” cancer suspicion some women reacted with relief while others did not. These greater changes that were observed in the group of women having false positives can only be regarded as harm arising from screening.
13. The process of the PhD study

The Psychological Consequences Questionnaire (PCQ)

Negative part: 12 items
Positive part: 10 items

Translation and adaptation process including a bilingual panel and a lay panel

6 ambiguous items reformulated
1 ambiguous item reformulated

The Danish version of the PCQ

Negative part: 18 items
Positive part: 11 items

Focus group interviews assessing the content validity of the PCQ-DK18 including women who had had an abnormal screening mammography later confirmed to be false-positive

Abnormal screening mammography

15 new items generated to cover the area
All 18 items found to be relevant

33-item new questionnaire

False-positive screening mammography

3 items found to be irrelevant
The content of 8 items found to be relevant
5 new items generated to cover the area

Reformulated
13-item new questionnaire

Statistical test of the psychometric properties of the PCQ and the Nottingham Health Profile (NHP)

5 “poor” items identified

Consequences of Screening – in Breast Cancer (COS-BC)

COS-BC part 1 (28 items)
Six dimensions:
- anxiety (6)
- behavioural (7)
- sense of dejection (6)
- sleep (2)
- breast examination (2)
- sexuality (2)
Three single items

COS-BC part 2 (13 items)
Four dimensions:
- existential values (6)
- social network (3)
- feeling relax/calm (2)
- anxiety about/belief in (not) having breast cancer (2)
No single items

2 sleep-items from NHP suggested to be added to part 1
14. Conclusions and implications for practice and research

14.1. Conclusion

This PhD study began with a literature review concluding that generic instruments are inadequate for measuring psychosocial consequences of cancer screening. The most appropriate questionnaire identified for assessment of short-term psychosocial consequences of false-positive screening mammography was the Psychological Consequences Questionnaire (PCQ). However, little evidence was found to suggest that this, or any other instrument, could adequately detect long-term psychosocial consequences of screening mammography.

After translating and adapting the PCQ into Danish the instrument was tested for content validity in a setting of abnormal and false-positive screening mammography. It was found that the PCQ captured important negative psychosocial consequences of abnormal screening mammography. However, the measure lacked content coverage for women receiving abnormal and false-positive screening results.

Fifteen new items were added to the Danish version of the PCQ in order to broaden the spectrum of the psychosocial consequences of abnormal screening mammography covered by the measure. The additional items stem entirely from the qualitative findings of the study. This process resulted in a new draft questionnaire containing 33 items. The psychometric properties of the 33-item measure were explored in groups of women having normal, abnormal and false-positive screening mammography and women diagnosed with breast cancer.

The statistical analyses (using international gold standards) revealed 27 “good” items measuring different attributes of the same overall construct; the psychosocial consequences of abnormal screening mammography. Six subscales (including altogether 25 items) covered anxiety, behaviour, dejection, sexuality, sleep and breast examination. Two single items covered feeling less attractive and keeping busy to take mind off things.

The instrument also included an additional single item measuring sick leave.

The 33-item questionnaire was not relevant and did not cover the specific psychosocial consequences occurring after final false-positive diagnosis. To cover these specific psychosocial consequences a draft version of a new measure containing 13 items was developed.

The statistical analyses (again using international gold standards) exploring the psychometric properties of this 13-item questionnaire revealed four subscales. The subscales covered the impact on the women’s existential values, relationship with their social network, being less or more relaxed/calm, and less or more anxious about breast cancer/belief in having or not-having breast cancer.

As a result of this work preliminary evidence for a valid and reliable condition-specific instrument measuring psychosocial consequences of abnormal and false-positive screening mammography has been established. This valid condition-specific questionnaire consists of two parts, measuring;

1. psychosocial consequences of abnormal screening mammography (part 1)
2. long-term psychosocial consequences of false-positive screening mammography (part 2)

The instrument is entitled Consequences of Screening on Breast Cancer (COS-BC).
The PhD study shows that none of the quantitative studies of the psychosocial consequences of screening mammography conducted previously, have used adequate questionnaires. Although using inadequate measures, all of these studies reported consistently negative short-term psychosocial consequences of false-positive screening mammography. In contrast, the results of studies of long-term psychosocial consequences were ambiguous. This is most likely due to the inadequacy of the measures used.

Given the inadequacy of the measure used to assess long-term psychosocial consequences, the results of previously conducted studies do not provide conclusive evidence of such consequences. Hence, the conclusions drawn from these studies regarding long-term psychosocial consequences of false-positive results of screening mammography should remain tentative.

This PhD study shows that having an abnormal screening mammography causes significant negative psychosocial consequences. Two weeks after abnormal screening mammography the negative consequences decrease. However, they are still of measurable significance. The study also shows that women’s attitudes towards existential values change significantly from one to six months after a false-positive screening mammography. Women also experienced an impact on the relationship to their social network, they felt more or less relaxed and calm and, finally, they were either more or less confident/anxious about having breast cancer.

The changes after the false-positive result can only be regarded as effects of a fear induced by the abnormal screening mammography and later the relief induced by the final false-positive result. It is difficult to regard these changes as anything but harm caused by the screening programme.

14.2. Implications for practice and research

14.2.1. Implications for practice
This study shows that there are substantial negative psychosocial consequences associated with having an abnormal screening mammography later confirmed to be false-positive. Consequently, the number of women having false-positive screening mammography should be kept to a minimum. Letters and folders posted at invitation to screening should contain information on the negative psychosocial consequences arising from abnormal and false-positive screening results. This information should be based on the new understanding of psychosocial consequences of screening mammography revealed in the present study. This would make the information on benefits and harm of breast cancer screening more balanced and allow the women invited to make an informed choice.

14.2.2. Implications for research
The COS-BC has been translated into Dutch, English and Norwegian and is currently in use in a major Danish survey. Data from the Danish survey and from studies in other countries may provide valid new knowledge, especially about the long-term psychosocial consequences of false-positive screening mammography.
With an adequate instrument available it would be of interest if more studies on psychosocial consequences of breast cancer screening were conducted in countries running breast screening programmes. Countries that have not yet implemented breast cancer screening should conduct studies on psychosocial consequences of false-positive screening mammography prior to their implementation in order to fulfil the international criteria of screening.

Results from new surveys using the COS-BC may generate additional new understanding of the detailed measurement of psychosocial consequences of screening mammography. The results from such studies would contribute to the balance sheet of the benefits versus the harm of the screening programme. This can be used as a part of the decision-making process for the implementation, or non-implementation, of breast cancer screening.

There is a need for valid instruments with high content validity and good psychometric properties in the research area of psychosocial consequences of screening for any cancers. Some items of the COS-BC are specific for women participating in breast cancer screening. The relevance of the COS-BC, without the breast cancer specific items, is currently being explored in the setting of lung cancer screening. It is also planned to test the instrument in a setting of cervical cancer screening and hopefully in the future, in other cancer screening programmes.

The ultimate goal is to establish core-items and core-subscales relevant for any kind of cancer screening. Using the core-questionnaire, COS, specific items can be added, relevant to the specific cancer screening programme and generated from interviews with the target populations.

The core-questionnaire COS would make it possible to potentially compare the extent of the psychosocial consequences of false-positive screening results in different cancer screening programmes. Adding specific items relevant to each cancer screening programme to the COS would potentially make it possible to demonstrate the psychosocial consequences of false-positive results in a particular screening programme, for example in colorectal cancer screening.
15. Reference list


60. von Bulow B. [Psychological consequences of breast cancer screening among healthy women]. [Review] [34 refs] [Danish]. Ugeskrift for Laeger 2000;162:1053-9.


78. McKenna SP, Doward LC. The translation and cultural adaptation of patient-reported outcome measures. *Value Health* 2005;8:89-91.


93. Rimer BK, Bluman LG. The psychosocial consequences of mammography. [Review] [27 refs]. *Journal of the National Cancer Institute* 1997; Monographs.:131-8.


121. Pallant JF., Tennant A. An introduction to the Rasch measurement model: An example using the Hospital Anxiety and Depression Scale (HADS). *British Journal of Clinical Psychology* 2006;In press.


16. Abbreviations and professional terms

Carcinoma in situ, CIS: Carcinoma in situ is a condition in between being healthy and having a cancer.

CFA: Confirmatory factor analyses.

Condition-specific questionnaire: A condition-specific questionnaire measures a specific condition in a specific population in contrast to a generic questionnaire that measures overall concepts e.g. health status, anxiety, depression, psychiatric morbidity etc.

COS: An acronym of Consequences of Screening. COS is a core-questionnaire that encompasses core-items and core-subscases relevant for any kind of cancer screening.

COS-BC: An acronym of Consequences of Screening on Breast Cancer. The COS-BC is a condition-specific questionnaire consisting of two parts, measuring:
- psychosocial consequences of abnormal screening mammography (part 1)
- long-term psychosocial consequences of false-positive screening mammography (part 2)

CTT: Classical Test Theory.

Differential item functioning, DIF: Differential item functioning is where an item functions differently in subpopulations such as in an intervention group and a control group.

DNBH: The Danish National Board of Health.

EU: The European Union.

False negative screening result: A sick person having a normal screening test.

False positive: Abbreviation for “false-positive screening mammography”.

False positive screening result: A healthy person having an abnormal screening test result.

Focus group interviews: Group interviews where both the issue(s) discussed and the participants in the group are focused.

Generic questionnaire: A generic questionnaire measures overall concepts e.g. health status, anxiety, depression, psychiatric morbidity etc. in contrast to a condition-specific measures that is developed to measure a specific condition in a specific population.

GHQ: The General Health Questionnaire, a generic questionnaire designed as a screening tool of psychiatric disorders in non-psychiatric clinical settings such as primary care or general medical out-patients.
<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>HADS</td>
<td>The Hospital Anxiety and Depression Scale, a generic questionnaire developed to provide a reliable screening test for psychiatric disorders in patients in non-psychiatric hospital departments. The HADS has two subscales; a depression and an anxiety subscale.</td>
</tr>
<tr>
<td>Incidence screening round</td>
<td>The second or following screening rounds.</td>
</tr>
<tr>
<td>Interval cancer</td>
<td>A cancer that appears in between two screening rounds.</td>
</tr>
<tr>
<td>IRT</td>
<td>Item Response Theory.</td>
</tr>
<tr>
<td>Item</td>
<td>A question with its corresponding response options.</td>
</tr>
<tr>
<td>Latent variable, latent trait</td>
<td>A variable which is unobservable but is supposed to enter into the structure of a system under study, such as for example depression.</td>
</tr>
<tr>
<td>Long-term consequences</td>
<td>Consequences that occur more than three months after the event.</td>
</tr>
<tr>
<td>Mammography</td>
<td>X-ray examination of the breast.</td>
</tr>
<tr>
<td>Mastectomy</td>
<td>Surgical removal of the breast.</td>
</tr>
<tr>
<td>Meta-analysis</td>
<td>The summarisation of evidence on an issue by combined analysis of distinct studies.</td>
</tr>
<tr>
<td>Needle biopsy</td>
<td>Cells that have been taken out with a thin piece of metal.</td>
</tr>
<tr>
<td>NHP</td>
<td>The Nottingham Health Profile, a generic health status questionnaire.</td>
</tr>
<tr>
<td>Occult</td>
<td>For example a disease not accompanied by readily discernible signs or symptoms.</td>
</tr>
<tr>
<td>PCQ</td>
<td>The Psychological Consequences Questionnaire.</td>
</tr>
<tr>
<td>PCQ-DK</td>
<td>The Danish version of the Psychological Consequences Questionnaire.</td>
</tr>
<tr>
<td>PCQ-DK33</td>
<td>A 33-item Danish questionnaire with high content validity for women having an abnormal screening mammography.</td>
</tr>
<tr>
<td>Pilot study</td>
<td>A study, usually on a small scale, carried out prior to the main study, primarily to gain information to improve the efficiency of the main study.</td>
</tr>
<tr>
<td>PLISSIT model</td>
<td>A model that offers an approach to communicating with people about sexuality.</td>
</tr>
</tbody>
</table>
Positive predictive value
The positive predictive value of a screening test expresses how many persons with a positive screening result actually have the disease.

Prevalence
The prevalence of e.g. disease is the number of existing persons with the disease in a population at a designated time.

Prevalence screening round
The first screening round.

Psychometric properties
Psychometric is relating to the measurement of mental abilities and qualities. Psychometric properties are in this context an overarching term for the validity and reliability of a psychometric measure.

Randomised controlled trial
In a randomised trial persons are by lottery chosen to be put in either one or more intervention groups or a control group. In the control group persons are having for example placebo treatment or health care as usual. The persons in the intervention group are given the treatment or the intervention explored.

Response categories
The available possibilities of responses to a question in an item.

Response options
The different possibilities of responding to a question in an item.

Salutogenic
Salutogenesis is a view of health which suggests that health is a dynamic process between opposite positions of total wellness and total illness. The development of positive resources is required for good health. These resources are infinitely variable and unique to the individual. They may be personal such as high self-esteem and cognitive ability or external such as living in a safe environment and having a good social support.

Screening mammography
A screening mammography is one done in women who have no signs of breast cancer. It usually involves two X-rays of each breast.

Screening recall clinics
Persons having an abnormal screening test will have further examinations to find out if the screening test result is true or false. These further examination are conducted in so called recall clinics.

Sensitivity
A screening test’s ability to find persons with disease.

SF-36
Short-form-36, a generic health status questionnaire.

Short-term consequences
Consequences that occur within three months after the event.
Specificity

A screening test’s ability to identify healthy persons.

STAI

The State-Trait Anxiety Inventory, a generic questionnaire measuring both state and trait anxiety in normal adults.

Validity, concurrent

Contains convergent validity and divergent validity, for example when a new measure is correlated to a existing valid measure the correlation between the measure can converge and diverge.

Validity, construct

The overarching concept of validity. Almost any kind of information about a test can contribute to an understanding of its construct validity. Fundamentally, all validation is construct validation, in the sense that all validity evidence contributes to (or undercuts) the empirical grounding or trustworthiness of the score interpretation”.

Validity, content

Content validity encompasses content relevance and content coverage. If questions, response categories and/or items are relevant for the target population they are said to have content relevance. If questions, response categories and/or items are covering the area explored they are said to have content coverage.

Validity, criterion

Concurrent validity and predictive validity are unified into the term criterion validity. For further explanation see those kinds of validity.

Validity, known group

Also called extreme groups validity. A test where the measure is given to two groups; one of which has the trait or behaviour, and the other which does not. The former group should score significantly higher (or lower) on the instrument.

Validity, predictive

An example of predictive validity could be a test’s ability to predict if students tested before they were admitted to university whether or not they would graduate three years later as bachelors.

Concurrent validity and predictive validity are unified into the term criterion validity. For further explanation see those kinds of validity.

WHO

World Health Organisation.
17. Appendix I - the Psychological Consequences Questionnaire (PCQ)

Part 1, negative psychological consequences of the actual act of participation in breast cancer screening:

We would like to find out about women’s experiences of the Breast X-ray Program. We would therefore like you to answer the questions on this questionnaire as best you can:

*Over the last week* how often have you experienced the following things because of *thoughts and feelings about breast cancer*?

1. Had trouble sleeping (P)
2. Experienced a change in appetite (P)
3. Been unhappy or depressed (E)
4. Been scared and panicky (E)
5. Felt nervous or strung up (E)
6. Felt under strain (E)
7. Found you have been keeping things from those who are close to you (S)
8. Found yourself taking things out on other people (S)
9. Found yourself noticeable withdrawing from who are close to you (S)
10. Had difficulty doing things around the house which you normally do (P)
11. Had difficulty meeting work or other commitments (P)
12. Felt worried about your future (E)

**Response options:**

0) Not at all
1) Rarely
2) Some of the time
3) Quite a lot of the time

**Sub-scales:**

(E) Emotional
(P) Physical
(S) Social
Part 2, positive psychological consequences of the actual act of participation in breast cancer screening:

All things considered, would you say your experiences at the Breast X-ray Program have caused any of the following?

1. A sense of reassurance that you do not have breast cancer (E)
2. Feeling more relaxed (E)
3. Improved relationship with friends or relations (S)
4. Feeling more able to do things which I normally do (P)
5. Feeling more able to meet your home and/or work responsibilities (P)
6. Feeling more hopeful about the future (E)
7. Feeling less anxious about breast cancer (E)
8. Getting on better with those around you (S)
9. Been sleeping better (P)
10. A greater sense of well being (E)

Response options:

0) Not at all
1) A little bit
2) Quite a bit
3) A great deal

Sub-scales:

(E) Emotional
(P) Physical
(S) Social
18. **Appendix II - overall search strategy**

The overall search strategy for the database MEDLINE:

Grundlæggende beslutninger for del 1

Hanne & John, efter fokusgruppe 130202, i perioden 13. til 20. februar 2002:

Besvarelserne skal så vidt mulig være grader. Det er i langt de fleste tilfælde mere naturligt at besvare de eksisterende spørgsmål - og de nye spørgsmål - med grader og ikke frekvenser. Hvor det ikke er muligt at besvare spørgsmålene med grader bruges frekvenser.

Graderne bliver:
- Ja, meget
- Ja, noget
- Ja, lidt
- Nej Slet

Alternativt til den af sprogpanelet foreslåede frekvens-svarmulighederne (som blev godkendt af lægpanelet og ingen i fokusgruppen 130202 havde bemærkninger til), foreslås nedenstående svarmuligheder:
- Ja, mange gange
- Ja, nogle gange
- Ja, få gange
- Nej, slet ikke

Til fokusgruppen 200202 præsenteredes ovenstående svarmuligheder, både grader, og de to frekvens- svarmuligheder. Fokusgruppen 200202 valgte den nye foreslåede frekvens-svarmulighed.
We would like to find out about women’s experiences of the Breast X-ray Program. We would therefore like you to answer the questions on this questionnaire as best you can:

Del 1

Vi vil gerne vide hvilke oplevelser kvinder har i forbindelse med røntgenundersøgelse af brystet (mammografi). Derfor vil vi bede Dem besvare de følgende spørgsmål, så godt De kan. (sprogpanel)

For at blive bedre til vort arbejde, vil vi gerne vide hvilke oplevelser De har haft i forbindelse med røntgenundersøgelse af brystet (mammografi). Besvar derfor venligst følgende spørgsmål, så godt De kan. (lægpanel)

Gennem dette spørgeskema, håber vi på, at få at vide hvilke oplevelser De har haft i forbindelse med røntgenundersøgelse af brystet (mammografi). Besvar derfor venligst følgende spørgsmål. (H&J 180202)

Gennem dette spørgeskema, håber vi på, at få at vide hvordan kvinder, der bliver inviteret til en befolkningsundersøgelse, har det før og efter røntgenundersøgelse af brystet (mammografi). Besvar derfor venligst følgende spørgsmål uanset om De agter at få brystet undersøgt. (H&J 280202)
Over the last week how often have experienced the following things because of thoughts and feelings about breast cancer:

Hvor ofte har De i den sidste uges tid oplevet følgende på grund af følelser og tanker om brystkræft? (sprogpanel)

Del 1

Hvor ofte har De - i den sidste uges tid - oplevet følgende på grund af tanker om brystkræft? (lægpanel)


| De originale svarmuligheder fra PCQ | □ Not at all  
□ Rarely  
□ Some of the time  
□ Quite a lot of the time |
| Svarmuligheder oversat af sprogpanel og godkendt af lægpanel | □ Nej, slet ikke  
□ Ja, men sjældent  
□ Ja, ind imellem  
□ Ja, ofte |
| Svarmuligheder oversat af sprogpanel og godkendt af lægpanel med anden grafisk layout (H&J180202) | Ja, ofte □  
Ja, ind imellem □  
Ja, men sjældent □  
Nej, slet ikke □ |
| Svarmuligheder foreslået af HT og JB, alternativt til ovenstående frekvenser med anden grafisk layout og anden rækkefølge (180202). | Ja, mange gange □  
Ja, nogle gange □  
Ja, få gange □  
Nej, slet ikke □ |
| Nye svarmuligheder foreslået af HT og JB, der hvor ”grader” er mere forståelige og anvendelige en ”frekvenser”. Desuden er layout ændret og ligeså rækkefølgen (180202). | Ja, meget □  
Ja, noget □  
Ja, lidt □  
Nej, slet ikke □ |
Nyt design og de nye svarmuligheder som fokusgruppen 200202 fortræk frem for det tidligere design og den af sprogpanelet foreslåede svarmulighed. De valgte selv hvor man skulle svare med grader og hvor man skulle svare med frekvenser.

Har De - i den sidste uges tid - oplevet følgende på grund af tanker om brystkræft?

<table>
<thead>
<tr>
<th>Åndret pga. ovenstående generelle betragtninger 280202</th>
<th>Nej, slet ikke</th>
<th>Ja, lidt</th>
<th>Ja, noget</th>
<th>Ja, meget</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Nej, slet ikke</td>
<td>Ja, få gange</td>
<td>Ja, nogle gange</td>
<td>Ja, mange gange</td>
</tr>
</tbody>
</table>

1. Jeg har været længe om at falde i søvn. ☐ ☐ ☐ ☐ ☐

Fokusgruppe 130302: Hvis man ikke længere er seksuel aktiv, og dermed ikke har gjort sig tanker om nedenstående, mangler svarmuligheden ”Ved ikke” til disse to items.

<table>
<thead>
<tr>
<th>Nej, slet ikke</th>
<th>Ja, lidt</th>
<th>Ja, noget</th>
<th>Ja, meget ikke</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nej, slet ikke</td>
<td>Ja, lidt</td>
<td>Ja, noget</td>
<td>Ja, meget ikke</td>
</tr>
</tbody>
</table>

30. Jeg har haft mindre lyst til sex. ☐ ☐ ☐ ☐ ☐ ☐

31. Jeg har haft mindre lyst til at få kærtægnet mit bryst. ☐ ☐ ☐ ☐ ☐ ☐
1. Had trouble sleeping (P)  
(sprogpanel) Jeg har haft søvnproblemer.  
(lægpanel) Jeg har haft søvnproblemer.  

På båndet fra fokusgruppe130202, er der både problemer med at sove, og falde i søvn. Søvnproblemer opfattes som om man har svært ved at sove, hvorfor dem der har svært ved at falde i søvn, ikke nødvendigvis ville krydse af i andet end ”Nej slet ikke”. Derfor erstattes dette spørgsmål med to spørgsmål fra NHP’s søvnskala, spørgsmål 4 og 5. I fokusgruppen 200202 fortalte en kvinde desuden at hun havde måtte tage sovepiller for at sove, NHP 1, søvnskala:  
(H&J efter fokusgruppe130202) 1a – Jeg er længe om at falde i søvn.  
(H&J efter fokusgruppe130202) 1b – Jeg sover dårligt om natten.  
(H&J efter fokusgruppe200202) 1c - Jeg har taget piller for at sove.  
(H&J efter fokusgruppe270202) 1a – Jeg har været længe om at falde i søvn.  
(H&J efter fokusgruppe270202) 1b – Jeg har sovet dårligt om natten.  
(H&J efter fokusgruppe270202) 1c - Jeg har taget piller for at sove.  

2. Experienced a change in appetite (P)  
(sprogpanel) Min appetit har ændret sig.  
(lægpanel) Min appetit har ændret sig.  

3. Been unhappy or depressed (E)  
(sprogpanel) Jeg har været ked af det eller deprimeret.  
(lægpanel) 3a - Jeg har været ked af det.  
(lægpanel) 3b - Jeg har været deprimeret.  

4. Been scared and panicky (E)  
(sprogpanel) Jeg har været angst og har været ved at gå i panik.  
(sprogpanel) Jeg har været bange og har været ved at gå i panik.  
(lægpanel) 4a - Jeg har været bange for at have brystkræft  
(H&J – kvinder med brystkræft skal også udfylde skemaet og mere tro mod originalskemaet)  
4a - Jeg har været bange.  
(lægpanel) 4b – Jeg har været ved at gå i panik.  

5. Felt nervous or strung up (E)  
(sprogpanel) Jeg har været nervøs.  
(lægpanel) Jeg har været nervøs.
6. Felt under strain (E)
   (sprogpanel) Jeg har været stresset.
   (lægpanel) Jeg har oplevet hverdagen uoverkommeligt.
   (fokusgruppe130202, forståelsesproblem med ”uoverkommelig”, foreslås)
   (fokusgruppe130202) Det har været svært at komme gennem hverdagen.
   (fokusgruppe270202, forståelsesproblemer og overlap, foreslås)
   (fokusgruppe270202, og godkendt af fokusgruppe130302) Jeg har følt, at tiden har været lang.

7. Found you have been keeping things from those who are close to you (S)
   (sprogpanel) Jeg har holdt mine tanker for mig selv.
   (lægpanel) Jeg har holdt mine tanker for mig selv.

8. Found yourself taking things out on other people (S)
   (sprogpanel) Jeg har ladet det gå ud over andre.
   (lægpanel) Jeg har været opfarende.
   (fokusgruppe130202, forståelsesproblem med ordet ”opfarende”, hvorfor følgende foreslås. Der
   var enighed om at ”irritabel” og ”uligevægtig”, var to forskellige ting og ”pirrelig” og
   ”irritabel” lignede hinanden)
   (fokusgruppe130202, fortrukket af fokusgruppe200202) Jeg har været irritabel.
   (fokusgruppe130202, fravalgt af fokusgruppe200202) Jeg har været pirrelig.
   (fokusgruppe130202. Fokusgruppe270202, sprogligt problem med uligevægtig)
   Jeg har været uligevægtig.
   (fokusgruppe270202) 8b - Jeg har været ude af balance.

9. Found yourself noticeable withdrawing from who are close to you (S)
   (sprogpanel) Jeg har trukket mig væk fra mine nærmeste.
   (lægpanel, fravalgt af fokusgruppe200202) Jeg har trukket mig væk fra mine nærmeste.
   (fokusgruppe130202) Jeg har trukket mig ind i mig selv.
   (fokusgruppe130202) Jeg har været mere stille.
   (fokusgruppe,bånd130202, fravalgt af fokusgruppe200202) Jeg har været mere tavs.
   Fokusgruppe200202 blev præsenteret for de ovenstående spørgsmål og angav at de to øverste
   lignede hinanden og de fortrak nummer to og at de to nederste også lignede hinanden og de
   fortrak de øverste af disse to, således at dette item herefter er to items:
   (fokusgruppe200202) 9a - Jeg har trukket mig ind i mig selv.
   (fokusgruppe130202) 9b – Jeg har været mere stille.
   (fokusgruppe200202) 9b – Jeg har mere stille.
10. Had difficulty doing things around the house which you normally do (P)
(sprogpanel) Jeg har haft svært ved at klare det huslige arbejde.
(lægpanel) Jeg har haft svært ved at klare opgaver i hjemmet.

11. Had difficulty meeting work or other commitments (P)
(sprogpanel) Jeg har haft svært ved at klare mit arbejde eller lignende opgaver.
(lægpanel) Jeg har haft svært ved at klare mit arbejde eller andre lignende opgaver.

12. Felt worried about your future (E)
(sprogpanel) Jeg har været bekymret for min fremtid.
(lægpanel) Jeg har været bekymret for min fremtid.

Ekstra
(fokusgruppe130202) Jeg har haft hovedpine.

Ekstra
(fokusgruppe130202) Jeg har været træt.

Ekstra
(fokusgruppe130202) Jeg har været bekymret.

Ekstra
(fokusgruppe130202) Jeg har ryddet mere op i hjemmet.
(H&J280202 ændrede dette spørgsmål, fokusgruppe130202 foreslog selv noget i denne retning som supplerende items) Jeg har flygtet fra min tanker, ved at foretage mig noget praktisk.

Ekstra
(fokusgruppe,bånd130202) Jeg har undersøgt mit brystet mere.
(fokusgruppe,bånd200202) Jeg har undersøgt mit brystet.

Ekstra
(fokusgruppe,bånd130202) Jeg har betragtet mit bryst i spejlet mere.
(fokusgruppe,bånd200202) Jeg har betragtet mit bryst i spejlet.

Ekstra
(fokusgruppe,bånd130202) Jeg har haft flere symptomer (prikken, stikken, smerter) fra mit bryst.
(fokusgruppe,bånd200202) Jeg har haft symptomer (prikken, stikken, smerter) fra mit bryst.

Ekstra
(fokusgruppe,bånd130202) Jeg har haft svært ved at koncentrere mig.

Ekstra
(fokusgruppe,bånd130202) Jeg har været urolig.
Ekstra

(fokusgruppe,bånd130202) Jeg har været sygemeldt.
(fokusgruppe200202) Hvor mange sygedage har De haft i den sidste uges tid?
0 dage, 1-2 dage, 3-4 dage, 5 eller mere

Ekstra

(fokusgruppe,bånd130202) Jeg har været rastløs.

Ekstra

(fokusgruppe130302) Jeg har følt mig mindre attraktiv som kvinde.

Ekstra

(fokusgruppe200302) Jeg har haft mindre lyst til sex, end jeg plejer.
(fokusgruppe270302) Jeg har haft mindre lyst til sex.

Ekstra

(fokusgruppe200302) Jeg har haft mindre lyst til at få kærtægnet mit bryst.

Ekstra

(fokusgruppe260902) Jeg har følt mig handlingslammet.

Subscales:

(E) Emotionel
(P) Physical
(S) Social
Grundlæggende beslutninger for del 2

Hanne & John, efter fokusgruppe130202 i perioden 13. til 20. februar 2002:

Svarmulighederne foreslået af fokusgruppen130202 virker ikke. Man kan ikke kun have ”ja, nej og det samme”. De originale virker slet ikke. Det er ændringer vi er interesseret i at måle, hvorfor der bør være flere svarmuligheder. I Health Measuring Questionaire er der ni svarmuligheder, fire negative, en neutral (det samme som før) og fire positive. Vi har valgt at have tre grader/frekvens i første del af dette spørgeskema, hvorfor det ville være logisk at have syv svarmuligheder i del to, nemlig tre negative, et neutalt og tre positive. Vi vælger derfor også at ændre grafikken radikalt.

I de originale svarmuligheder er der et problem, da nogle af spørgsmålene er identiske eller sammenfaldende i de to dele. Det gælder disse spørgsmål angivet med originale numre:

Del 1, spørgsmål 6 og del 2 spørgsmål 4
Del 1, spørgsmål 10 og del 2 spørgsmål 5
Del 1, spørgsmål 11 og del 2 spørgsmål 5
Del 1, spørgsmål 1 og del 2 spørgsmål 9

Hvis man forestiller sig, at en kvinde ved baseline-målingen svarer at hun ingen søvnproblemer har. Derefter svarer hun i udredningsfasen som falsk positiv at hun har mange søvnproblemer. En måned efter svarer hun at hun har nogen eller lidt søvnproblemer. Samtidig skal hun nu for første gang besvare del 2 og her krydser hun måske af, at hun har fået lidt mindre søvnproblemer end før undersøgelser. Årsagen kan være at hun har glemt hvordan hun havde det ved baseline, men krydser af i forhold til den bedring hun har oplevet fra udredningsfasen til nu.

Der er således klare problemer med ovenstående sammenfaldende spørgsmål i del 1 og 2. Desuden er det vi ønsker at måle som ændringer i ”positive” eller ”negativ” retning ikke små konkrete hændelser, men større diffuse eksistentielle ændringer. Det er også sådanne spørgsmål kvinderne selv foreslår i fokusgruppen130202.

Med ovenstående to argumenter, besluttes det at sløjfe spørgsmålene 4,5 og 9, fra det originale spørgeskemmas del 2. Det giver i den oversatte form, der er godkendt af lægpanelet, spørgsmålene 4,5,6 og 10.
All things considered, would you say your experiences at the Breast X-ray Program have caused any of the following:

**Del 2**

Har Deres oplevelser med røntgenundersøgelse af brystet, alt taget i betragtning, medført noget af følgende: (sprogpanel)

**Del 2**

Har Deres oplevelser efter røntgen-undersøgelsen af brystet, alt taget i betragtning, medført følgende: (lægpanel)

**Del 2**

Har Deres oplevelser efter undersøgelsene af brystet er afsluttet, alt taget i betragtning, medført følgende: (fokusgruppe130202)

(fokusgruppe260902) Sæn kun ét kryds ved hvert spørgsmål

<table>
<thead>
<tr>
<th>De originale svarmuligheder fra PCQ</th>
<th>Not at all</th>
<th>A little bit</th>
<th>Quite a bit</th>
<th>A great deal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oversatte svar muligheder fra sprogpanel og godkendt af lægpanel</td>
<td>Nej, slet ikke</td>
<td>Ja, lidt</td>
<td>Ja, noget</td>
<td>Ja, meget</td>
</tr>
<tr>
<td>Svarmuligheder forslået af fokusgruppe130202</td>
<td>Nej</td>
<td>Det samme som før</td>
<td>Ja</td>
<td></td>
</tr>
<tr>
<td>Svarmuligheder og grafik foreslået af H&amp;J:</td>
<td>Meget mindre</td>
<td>Noget mindre</td>
<td>Lidt mindre</td>
<td>Det samme som før</td>
</tr>
</tbody>
</table>

Mindre

Mere
1. 

A sense of reassurance that you do not have breast cancer (E)

(sprogpanel) Jeg har tillid til, at jeg ikke har brystkræft.

(lægpanel) Jeg stoler på, at jeg ikke har brystkræft.

Da brystkræft patienter også skal besvare spørgeskemaet, må dette spørgsmål og spørgsmål 7 (begge del 2) fjernes fra de spørgeskemaer som udsendes til kvinder med sandt positiv svar.

Fokusgrupperne 200202 og 270203 fandt dette og spørgsmål 7 relevante. (Efter beslutningen om nyt grafisk design med 7 svarmuligheder, kan ”stole på” ikke fastholdes, hvorfor H&J180202 foreslår at det erstattes med ”tillid”):

Efter undersøgelse har jeg tillid til, at jeg ikke har brystkræft.

Fokusgruppe 130302 fandt dette spørgsmål relevant, men påpegede at man skal passe på når der står et ”ikke” i spørgsmålet. H&J130302 foreslår derfor:

(fokusgruppe200202) Efter undersøgelse er min tillid til, at jeg ikke har brystkræft.

2. 

Feeling more relaxed (E)

(sprogpanel) Jeg er mere afslappet

(lægpanel) Jeg er mere afslappet end før røntgen-undersøgelsen

(fokusgruppe130202) Jeg er mere afslappet end før undersøgelserne

(fokusgruppe200202) Efter undersøgelserne føler jeg mig: og så afslappet i svarmulighederne

3. 

Improved relationship with friends or relations (S)

(sprogpanel) Mit forhold til venner og familie er blevet bedre

(sprogpanel) Jeg har fået det bedre med mine venner eller familie

(lægpanel) Mit forhold til familie og venner er blevet bedre

3a (fokusgruppe130202) Mit forhold til familien er blevet tættere end før undersøgelserne

3b (fokusgruppe130202) Mit forhold til vennene er blevet tættere end før undersøgelserne

Er disse to spørgsmål sammenfaldene med spørgsmål 8? Skal diskuteres i fokusgruppe200202.

3a (fokusgruppe200202) Efter undersøgelserne er mit forhold til min familie:

3b (fokusgruppe200202) Efter undersøgelserne er mit forhold til vennerne:

4. 

Feeling more able to do things which I normally do (P)

(sprogpanel) Jeg har lettere ved at gøre det, jeg plejer

(sprogpanel) Jeg kan bedre klare det, jeg plejer

(lægpanel) Hverdagen er blevet mere overkommelig

(fokusgruppe130202) Det er lettere at komme gennem hverdagen end før undersøgelserne

UDGÅR
5. Feeling more able to meet your home and/or work responsibilities (P) 
   (sprogpanel) Det er lettere for mig at klare opgaver i hjemmet og/eller på arbejde 
   5a – Det er blevet lettere for mig at klare opgaver i hjemmet 
   (lægpanel) Det er blevet lettere for mig at klare mit arbejde eller andre lignende opgaver 
   (fokusgruppe130202) Det er blevet lettere for mig at klare mit arbejde eller andre lignende 
   opgaver end før undersøgelserne 

6. Feeling more hopeful about the future (E) 
   (sprogpanel) Jeg ser lysere på fremtiden 
   (lægpanel) Jeg ser lysere på fremtiden 
   (fokusgruppe130202) Jeg ser lysere på fremtiden end før undersøgelserne 
   (fokusgruppe200202) Efter undersøgelserne er mit syn på fremtiden blevet: og så ”lys” i 
   svarmulighederne 

7. Feeling less anxious about breast cancer (E) 
   (sprogpanel) Jeg er mindre bekymret for brystkræft 
   (lægpanel) Jeg er blevet mindre bekymret for brystkræft 
   (H&J, efter nyt design og svarmuligheder) Efter undersøgelserne er jeg er blevet mindre bekymret 
   for brystkræft: 
   (fokusgruppe200202) Efter undersøgelserne er min bekymring for brystkræft: 

8. Getting on better with those around you (S) 
   (sprogpanel) Jeg har det bedre med andre mennesker 
   (lægpanel) Jeg har fået det bedre med andre mennesker 
   (fokusgruppe200202) Efter undersøgelserne er mit forhold til andre mennesker: 
   Se spørgsmål 3a og 3b. 

9. Been sleeping better (P) 
   (sprogpanel) Jeg sover bedre 
   (lægpanel) Jeg sover bedre 
   (fokusgruppe130202) Jeg sover bedre end før undersøgelserne 

10. A greater sense of well being (E) 
    (sprogpanel) Jeg har en større følelse af velvære end før undersøgelserne 
    (lægpanel) Jeg har fået en større følelse af velvære 
    (fokusgruppe130202) Jeg har fået en større følelse af velvære end før undersøgelserne 
    (fokusgruppe200202) Efter undersøgelserne er min følelse af velvære:
(fokusgruppe130202) Jeg har fået mere ro end før undersøgelserne
(fokusgruppe260902) Efter undersøgelserne er min fornemmelse af indre ro:

(fokusgruppe130202) Jeg er blevet mere glad for livet end før undersøgelserne
(fokusgruppe200202) Efter undersøgelserne er min glæde ved livet:

(fokusgruppe130202) Jeg tænker mere over livet end før undersøgelserne
(fokusgruppe200202) Efter undersøgelserne har jeg tænkt over livet:

(fokusgruppe130202) Jeg sætter mere pris på livet end før undersøgelserne
(fokusgruppe200202) Efter undersøgelserne værdsætter jeg livet:

(fokusgruppe130202) Jeg er mere bevidst om at leve end før undersøgelserne
(fokusgruppe260902) Efter undersøgelserne er jeg bevidst om at leve

Subscales:

(E) Emotionel
(P) Physical
(S) Social
Appendix IV - the PCQ-DK33 (Danish version)
Spørgeskema
til kvinder, der har fået
brystet røntgen-undersøgt (mammografi)

Gennem dette spørgeskema, håber vi på at få at vide, hvordan kvinder har det før og efter røntgen-undersøgelse af brystet (mammografi).

Det er vigtigt, at De svarer på alle spørgsmålene.
Har De - **in den sidste uges tid** - oplevet følgende på grund af tanker om brystkræft?

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>Nej, slet ikke</th>
<th>Ja, lidt</th>
<th>Ja, noget</th>
<th>Ja, meget</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Jeg har følt mig mindre attraktiv som kvinde.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Jeg har været bekymret.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Jeg har været bekymret for min fremtid.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Jeg har været bange.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Jeg har været irriteret.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Jeg har holdt mine tanker for mig selv.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Jeg har sovet dårligt om natten.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. Jeg har flygtet fra min tanker, ved at foretage mig noget praktisk.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11. Jeg har haft svært ved at koncentrere mig.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12. Jeg har følt, at tiden har været lang.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Har De - i den sidste uges tid - oplevet følgende på grund af tanker om brystkæft?

<table>
<thead>
<tr>
<th></th>
<th>Nej, slet ikke</th>
<th>Ja, lidt</th>
<th>Ja, noget</th>
<th>Ja, meget</th>
</tr>
</thead>
<tbody>
<tr>
<td>15. Jeg har været ude af balance.</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>17. Jeg har været rastløs.</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>18. Jeg har været nervøs.</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>19. Jeg har været urolig.</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>20. Jeg har været længe om at falde i søvn.</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>21. Jeg har betragtet mit bryst i spejlet.</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>22. Jeg har trukket mig ind i mig selv.</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>23. Jeg har følt mig handlingslammet.</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>24. Jeg har været deprimert.</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>25. Jeg har haft symptomer (prikken, stikken, smerter) fra mit bryst.</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
</tbody>
</table>
Har De - **i den sidste uges tid** - oplevet følgende på grund af tanker om brystkræft?

<table>
<thead>
<tr>
<th></th>
<th>Nej, slet ikke</th>
<th>Ja, få gange</th>
<th>Ja, nogle gange</th>
<th>Ja, mange gange</th>
</tr>
</thead>
<tbody>
<tr>
<td>26. Jeg har haft svært ved at klare mit arbejde eller andre lignende opgaver.</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
</tr>
<tr>
<td>27. Jeg har haft hovedpine.</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
</tr>
<tr>
<td>28. Jeg har haft svært ved at klare opgaver i hjemmet.</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
</tr>
<tr>
<td>29. Jeg har været ved at gå i panik.</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
</tr>
<tr>
<td>30. Jeg har taget piller for at sove.</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Nej, slet ikke</th>
<th>Ja, lidt</th>
<th>Ja, noget</th>
<th>Ja, meget</th>
<th>Ved ikke</th>
</tr>
</thead>
<tbody>
<tr>
<td>31. Jeg har haft mindre lyst til sex.</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
</tr>
<tr>
<td>32. Jeg har haft mindre lyst til at få kærtegnet mit bryst.</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
</tr>
</tbody>
</table>

33. Hvor mange sygedage har De haft **i den sidste uges tid**?

<table>
<thead>
<tr>
<th></th>
<th>0 dage</th>
<th>1 - 2 dage</th>
<th>3 - 4 dage</th>
<th>5 dage eller mere</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
</tr>
</tbody>
</table>

**Tak for hjælpen**
21. Appendix V - the PCQ-DK33 (English version)

All the items in the English version of the PCQ-DK33 has been translated and adapted using a bilingual panel and a lay panel. However, the five identified “poor” items (no’s 7, 9, 25, 27 and 30) have not been translated using the two panels but only been ad hoc translated for this thesis. The English version of the questionnaire has not been field tested and statistically validated.
Questionnaire
for women
who have undergone
breast screening

With this questionnaire we hope to find out how women feel both before and after breast screening.

It is important that you answer all questions.
To what extent - over the last week - have you experienced the following because of thoughts about breast cancer?

<table>
<thead>
<tr>
<th></th>
<th>Not at all</th>
<th>A bit</th>
<th>Quite a bit</th>
<th>A lot</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. I have felt less attractive.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. I have been worried.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. I have been worried about my future.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. I have felt scared.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. I have been irritable.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. I have been quieter than normal.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Found you have been keeping things from those who are close to you</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. I have slept badly.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. I have been tired.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. I have kept busy to take my mind off things.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11. I have found it hard to concentrate.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12. I have felt time passed slowly.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13. My appetite has changed.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
To what extent - over the last week - have you experienced the following because of thoughts about breast cancer?

<table>
<thead>
<tr>
<th></th>
<th>Not at all</th>
<th>A bit</th>
<th>Quite a bit</th>
<th>A lot</th>
</tr>
</thead>
<tbody>
<tr>
<td>14. I have felt sad.</td>
<td></td>
<td>⬜</td>
<td>⬜</td>
<td>⬜</td>
</tr>
<tr>
<td>15. I have been upset.</td>
<td></td>
<td>⬜</td>
<td>⬜</td>
<td>⬜</td>
</tr>
<tr>
<td>16. I have examined my breasts.</td>
<td></td>
<td>⬜</td>
<td>⬜</td>
<td>⬜</td>
</tr>
<tr>
<td>17. I have felt restless.</td>
<td></td>
<td>⬜</td>
<td>⬜</td>
<td>⬜</td>
</tr>
<tr>
<td>18. I have been nervous.</td>
<td></td>
<td>⬜</td>
<td>⬜</td>
<td>⬜</td>
</tr>
<tr>
<td>19. I have been uneasy.</td>
<td></td>
<td>⬜</td>
<td>⬜</td>
<td>⬜</td>
</tr>
<tr>
<td>20. It has taken me a long time to fall asleep.</td>
<td></td>
<td>⬜</td>
<td>⬜</td>
<td>⬜</td>
</tr>
<tr>
<td>21. I have examined my breasts in the mirror.</td>
<td></td>
<td>⬜</td>
<td>⬜</td>
<td>⬜</td>
</tr>
<tr>
<td>22. I have withdrawn into myself.</td>
<td></td>
<td>⬜</td>
<td>⬜</td>
<td>⬜</td>
</tr>
<tr>
<td>22. I have felt unable to cope.</td>
<td></td>
<td>⬜</td>
<td>⬜</td>
<td>⬜</td>
</tr>
<tr>
<td>23. I have been depressed.</td>
<td></td>
<td>⬜</td>
<td>⬜</td>
<td>⬜</td>
</tr>
<tr>
<td>25. I have had symptoms from my breasts (pins and needles).</td>
<td></td>
<td>⬜</td>
<td>⬜</td>
<td>⬜</td>
</tr>
</tbody>
</table>

Please turn over
To what extent - over the last week - have you experienced the following because of thoughts about breast cancer?

<table>
<thead>
<tr>
<th>Question</th>
<th>Not at all</th>
<th>A bit</th>
<th>Quite a bit</th>
<th>A lot</th>
</tr>
</thead>
<tbody>
<tr>
<td>26. I have had difficulty dealing with my work or other commitments.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27. I have had headache.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28. I have had difficulty doing everyday things around the house.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29. I have felt terrified.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30. I have taken pills to sleep.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>31. I have felt less interest in sex.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>32. I have not felt like having my breast caressed.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>33. How many days sick leave have you had during the last week?</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 days</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 - 2 days</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 - 4 days</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 or more days</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Thank you for your help.
22. Appendix VI – part 2 of the measure (Danish version)
Spørgeskema
til kvinder, der har fået
brystet røntgen-undersøgt (mammografi)

Gennem dette spørgeskema, håber vi på at få at vide, hvordan kvinder har det efter røntgen-undersøgelse af brystet (mammografi).

Det er vigtigt, at De svarer på alle spørgsmålene.
Har Deres oplevelser efter undersøgelserne af brystet er afsluttet, alt taget i betragtning, medført følgende:

Sæt kun ét kryds ved hvert spørgsmål

1. Efter undersøgelserne er min tillid til, at jeg ikke har brystkræft:

<table>
<thead>
<tr>
<th>Meget mindre</th>
<th>Noget mindre</th>
<th>Lidt mindre</th>
<th>Den samme som før</th>
</tr>
</thead>
<tbody>
<tr>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
</tbody>
</table>

Mindre tillid

<table>
<thead>
<tr>
<th>Lidt større</th>
<th>Noget større</th>
<th>Meget større</th>
</tr>
</thead>
<tbody>
<tr>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
</tbody>
</table>

Større tillid

2. Efter undersøgelserne er min bekymring for brystkræft:

<table>
<thead>
<tr>
<th>Meget større</th>
<th>Noget større</th>
<th>Lidt større</th>
<th>Den samme som før</th>
</tr>
</thead>
<tbody>
<tr>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
</tbody>
</table>

Større bekymring

<table>
<thead>
<tr>
<th>Lidt mindre</th>
<th>Noget mindre</th>
<th>Meget mindre</th>
</tr>
</thead>
<tbody>
<tr>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
</tbody>
</table>

Mindre bekymring

3. Efter undersøgelserne føler jeg mig:

<table>
<thead>
<tr>
<th>Meget mindre afslappet</th>
<th>Noget mindre afslappet</th>
<th>Lidt mindre afslappet</th>
<th>Det samme som før</th>
</tr>
</thead>
<tbody>
<tr>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
</tbody>
</table>

Mindre afslappet

<table>
<thead>
<tr>
<th>Lidt mere afslappet</th>
<th>Noget mere afslappet</th>
<th>Meget mere afslappet</th>
</tr>
</thead>
<tbody>
<tr>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
</tbody>
</table>

Mere afslappet

4. Efter undersøgelserne er min fornemmelse af indre ro:

<table>
<thead>
<tr>
<th>Meget mindre</th>
<th>Noget mindre</th>
<th>Lidt mindre</th>
<th>Den samme som før</th>
</tr>
</thead>
<tbody>
<tr>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
</tbody>
</table>

Mindre ro

<table>
<thead>
<tr>
<th>Lidt større</th>
<th>Noget større</th>
<th>Meget større</th>
</tr>
</thead>
<tbody>
<tr>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
</tbody>
</table>

Mere ro
Har Deres oplevelser efter undersøgelserne af brystet er afsluttet, alt taget i betragtning, medført følgende:

Sæt kun ét kryds ved hvert spørgsmål

5. Efter undersøgelserne er mit forhold til min familie:

- □ Meget mindre tæt
- □ Noget mindre tæt
- □ Lidt mindre tæt
- □ Det samme som før
- □ Lidt mere tæt
- □ Noget mere tæt
- □ Meget mere tæt

Mindre tæt | Mere tæt

6. Efter undersøgelserne er mit forhold til vennerne:

- □ Meget mindre tæt
- □ Noget mindre tæt
- □ Lidt mindre tæt
- □ Det samme som før
- □ Lidt mere tæt
- □ Noget mere tæt
- □ Meget mere tæt

Mindre tæt | Mere tæt

7. Efter undersøgelserne er mit forhold til andre mennesker:

- □ Meget dårligere
- □ Noget dårligere
- □ Lidt dårligere
- □ Det samme som før
- □ Lidt bedre
- □ Noget bedre
- □ Meget bedre

Dårligere | Bedre

8. Efter undersøgelserne er mit syn på fremtiden blevet:

- □ Meget mindre lyst
- □ Noget mindre lyst
- □ Lidt mindre lyst
- □ Det samme som før
- □ Lidt mere lyst
- □ Noget mere lyst
- □ Meget mere lyst

Mindre lys fremtid | Mere lys fremtid

Vend
Har Deres oplevelser **efter undersøgelserne** af brystet er afsluttet, alt taget i betragtning, medført følgende:

<table>
<thead>
<tr>
<th>Spørgsmål</th>
<th>Oplevelse</th>
<th>Meget mindre</th>
<th>Noget mindre</th>
<th>Lidt mindre</th>
<th>Det samme som før</th>
<th>Lidt større</th>
<th>Noget større</th>
<th>Meget større</th>
</tr>
</thead>
<tbody>
<tr>
<td>11. Efter undersøgelserne er jeg bevidst om at leve:</td>
<td>Mindre bevidst om at leve</td>
<td>[ ] Meget mindre</td>
<td>[ ] Noget mindre</td>
<td>[ ] Lidt mindre</td>
<td>[ ] Det samme som før</td>
<td>[ ] Lidt mere</td>
<td>[ ] Noget mere</td>
<td>[ ] Meget mere</td>
</tr>
<tr>
<td>12. Efter undersøgelserne har jeg tænkt over livet:</td>
<td>Færre tanker over livet</td>
<td>[ ] Meget mindre</td>
<td>[ ] Noget mindre</td>
<td>[ ] Lidt mindre</td>
<td>[ ] Det samme som før</td>
<td>[ ] Lidt mere</td>
<td>[ ] Noget mere</td>
<td>[ ] Meget mere</td>
</tr>
</tbody>
</table>

**Tak for hjælpen**
23. Appendix VII – part 2 of the measure (English version)

All the items in the English version of part 2 of the measure has been translated and adapted using a bilingual panel and a lay panel. However, the translation and adaptation was only conducted with five response categories. Therefore, the two response categories at each side of the response option “the same as before” are ad hoc translated for this thesis. The English version of the questionnaire has not been field tested and statistically validated.
Questionnaire
for women
who have undergone
breast screening

With this questionnaire we hope to find out how women feel after breast screening.

It is important that you answer all questions.
Taking everything into account, has your experience of the Breast Screening Programme caused any of the following:

**Please tick only one box for each question**

<table>
<thead>
<tr>
<th>Question</th>
<th>Options</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. After the examinations my belief that I do not have breast cancer is:</td>
<td>Much less</td>
</tr>
<tr>
<td>2. After the examinations my anxiety about breast cancer is:</td>
<td>Much greater</td>
</tr>
<tr>
<td>3. After the examinations I feel:</td>
<td>Much less relaxed</td>
</tr>
<tr>
<td>4. After the examinations I feel:</td>
<td>Much less calmer</td>
</tr>
</tbody>
</table>
Taking everything into account, has your experience of the Breast Screening Programme caused any of the following:

Please tick only one box for each question

<table>
<thead>
<tr>
<th>Question</th>
<th>Choices</th>
</tr>
</thead>
<tbody>
<tr>
<td>5. After the examinations my relationship with my family is:</td>
<td>Much less close</td>
</tr>
<tr>
<td></td>
<td>Less close</td>
</tr>
<tr>
<td>6. After the examinations my relationship with friends is:</td>
<td>Much less close</td>
</tr>
<tr>
<td></td>
<td>Less close</td>
</tr>
<tr>
<td>7. After the examinations my relationship with other people is:</td>
<td>Much worse</td>
</tr>
<tr>
<td></td>
<td>Worse</td>
</tr>
<tr>
<td>8. After the examinations my thoughts about the future are:</td>
<td>Much more pessimistic</td>
</tr>
<tr>
<td></td>
<td>More pessimistic</td>
</tr>
</tbody>
</table>

Please turn over
Taking everything into account, has your experience of the Breast Screening Programme caused any of the following:

Please tick only one box for each question

9. After the examinations my enjoyment of life is:

- [ ] Much less
- [ ] Less
- [ ] A little less
- [ ] The same as before
- [ ] A little more
- [ ] More
- [ ] Much more

10. After the examinations I value life:

- [ ] Much less
- [ ] Less
- [ ] A little less
- [ ] The same as before
- [ ] A little more
- [ ] More
- [ ] Much more

11. After the examinations my awareness of life is:

- [ ] Much less
- [ ] Less
- [ ] A little less
- [ ] The same as before
- [ ] A little greater
- [ ] Greater
- [ ] Much greater

12. After the examinations I have thought about the broader aspects of life:

- [ ] Much less
- [ ] Less
- [ ] A little less
- [ ] The same as before
- [ ] A little more
- [ ] More
- [ ] Much more

13. After the examinations my sense of well-being is:

- [ ] Much less
- [ ] Less
- [ ] A little less
- [ ] The same as before
- [ ] A little greater
- [ ] Greater
- [ ] Much greater

Thank you for your help