Identifying intimate partner violence in different ethnic groups in primary care - a systematic review and secondary data analysis

(Alex) Hardip Sohal

A thesis submitted for the degree of
Doctor of Medicine (Research)
University of London
Barts and the London School of Medicine and Dentistry, Queen Mary University of London

April 2011

The work presented in this thesis is all my own.

Dr (Alex) Hardip Sohal

"We are inclined to believe those whom we do not know because they have never deceived us"

Samuel Johnson,
English author, critic and lexicographer, (1709 – 1784)

Identifying intimate partner violence in different ethnic groups in primary care

Abstract

Background

Intimate partner violence (IPV), including physical, sexual and emotional violence, causes short and long term ill-health. Brief questions that can identify women from different ethnic groups experiencing IPV who present in clinical settings are a pre-requisite for an appropriate response from health services to this substantial public health problem.

Aim: To examine the evidence for the validity of questions trying to identify IPV in different ethnic groups and to determine whether their validity varies between ethnic groups.

Methods

Design: A systematic review and the secondary data analysis of a cross-sectional survey of four questions (HARK) identifying IPV in a primary care sample.

Main outcome measures: Systematic review - for each set of index questions identified, diagnostic accuracy indices, correlation coefficients, reliability measures, validity evidence based on response processes and test content were analysed and interpreted.

Secondary data analysis - diagnostic indices for IPV and its dimensions in three ethnic groups were calculated for the four HARK questions combined and for the individual HARK questions.

Results

Systematic review – there is no evidence of questions valid for identifying IPV in specific ethnic groups, including white groups.

Secondary data analysis - the optimal HARK cut off score of ≥ 1 was unaffected by the participants' ethnicity. The diagnostic indices generated using the HARK cut off of ≥ 1 remained at a high level, in all three ethnic groups. There were no significant ethnic differences in the diagnostic indices of the four combined and individual HARK questions' ability at identifying either IPV or its dimensions.

Conclusion

From the systematic review and secondary data analysis, there is no evidence that questions' validity for identifying IPV varies significantly between different ethnic groups. The secondary data analysis does provide evidence that four questions (the HARK) can identify IPV in self-classified UK census categories of African-Caribbean, south Asian, and white groups.

Table of Contents

Abstract…….3
Table of contents……...5
Acknowledgements…….14
Glossary……...16
Outputs related to thesis……18

Chapter 1: Background……...20
1.1. Introduction and Chapter Overview……20

1.2. Intimate Partner Violence……24
1.2.1. Is IPV a priority that should be addressed by primary care?.....24
1.2.1.1. Prevalence……25
1.2.1.2. Health Impact……...26
1.2.1.3. IPV Interventions……31
1.2.1.4. Approaches to identifying IPV……32
1.2.1.4.1. Screening……..32
1.2.1.4.2. Enquiry……33
1.2.1.4.2.1. Routine enquiry……33
1.2.1.4.2.2. Selective enquiry……..33

1.3. Measuring Validity……39
1.3.1. Measurement……40
1.3.1.1. Questions and their evaluation……41
1.3.2. Classical diagnostic accuracy paradigm……...43
1.3.3. Validation paradigm……45
1.3.4. Categorical and dimensional models……46
1.3.5. Integrating research paradigms……48
1.3.6. A categorisation of validity evidence……...49

1.3.6.1. Category A: Validity evidence based on the consequences of testing……...52
1.3.6.2. Category B: Validity evidence based on relations to other variables……53
1.3.6.2.1. Criterion performance studies……..53
1.3.6.2.1.1. Diagnostic accuracy studies….55
1.3.6.2.1.2. Criterion correlation studies….59
1.3.6.2.2. Association studies……...60
1.3.6.2.2.1. Between index scores and external variables……..60
1.3.6.2.2.2. Convergent validity studies…..60
1.3.6.2.2.3. Divergent validity studies…….61
1.3.6.2.3. Known group comparisons……...61
1.3.6.3. Category C: Validity evidence based on internal structure…..62
1.3.6.3.1. What is internal consistency reliability?.....62
1.3.6.3.2. Internal consistency and classical test theory...63
1.3.6.3.3. When should internal consistency not be considered…….63
1.3.6.3.4. Internal consistency in relation to other reliability measures……...64
1.3.6.3.5. Statistical methods to calculate internal consistency reliability……...65
1.3.6.3.5.1. Item-total correlation method...65
1.3.6.3.5.2. Inter-item correlation method...65
1.3.6.4. Correlation and correlation coefficients……...66
1.3.6.4.1. Pearson's product moment correlation coefficient…….67
1.3.6.4.2. Kappa coefficient…….67
1.3.6.4.3. Point biserial correlation coefficient……68
1.3.6.4.4. Spearman's correlation coefficient……...68
1.3.6.5. Category D: Validity evidence based on response processes..68
1.3.6.6. Category E: Validity evidence based on test content……...69
1.3.6.6.1. Translation of questions……...71

1.4. Ethnicity……..73
1.4.1. What is an ethnic group?.....73
1.4.1.1. Ethnicity's relationship to race……74
1.4.1.2. The phrase "Race / ethnicity"……...76
1.4.1.3. Studying ethnicity……77
1.4.2. Rationale for collecting ethnicity data in health research……78
1.4.2.1. Exposing health inequalities……78
1.4.2.2. Improving health……..79
1.4.2.3. Responding to increasing ethnic diversity……80
1.4.3. Dangers of collecting ethnicity data in health research……81
1.4.3.1. Racism……..81
1.4.3.2. Arbitrary classification……82
1.4.3.3. Inadequate analysis……...83
1.4.4. Five criteria to assess the use of ethnicity data by papers……..……..84

1.5. IPV Research and Ethnicity……...86
1.5.1. Research questions……...……89
1.5.2. Aims and objectives……90
1.5.3. Outline of the thesis……..91

Chapter 2: Methods……..93
2.1. Overview……..93

2.2. Systematic review……...94
2.2.1. Data sources and search strategy……..94
2.2.2. Study selection……95
2.2.3. Data extraction……96
2.2.4. Analysis of primary data extracted……...96
2.2.5. Quality appraisal……...97

2.3. Secondary data analysis……99
2.3.1. Sample size……102

2.3.2. Analysis……..103
2.3.2.1. HARK's ability to identify IPV in different ethnic groups…103
2.3.2.2. Each individual HARK question's ability to identify IPV in different ethnic groups……104
2.3.2.3. Each individual HARK question's ability to identify dimensions of IPV (emotional and physical IPV) in different ethnic groups……..104

Chapter 3: Results……...106
3.1. Overview……106

3.2. Systematic review results……107
3.2.1. Five studies reporting exclusively diagnostic accuracy……134
3.2.1.1. Single Safety Question……...134
3.2.1.2. Slapped, Threatened or Thrown (STaT) ……135
3.2.1.3. Partner violence screen (PVS) – two studies……..136
3.2.1.4. HARK……139
3.2.1.5. Overview of five studies reporting exclusively diagnostic accuracy……..141
3.2.2. Eight studies reporting validation paradigm methods with diagnostic accuracy……..141
3.2.2.1. Three question Chinese Abuse Assessment Screen (AAS)…141
3.2.2.2. Portuguese AAS's anchor question……142
3.2.2.3. Ongoing Violence Assessment Tool (OVAT)……...143
3.2.2.4. HITS – two studies……145
3.2.2.5. Five non-graphic domestic violence (DV) questions…….147
3.2.2.6. Behavioural Risk Factor Surveillance Survey (BRFSS)……148
3.2.2.7. Women's Experience with Battering Scale (WEB)……...150
3.2.2.8. Overview of eight studies reporting validation paradigm methods with diagnostic accuracy……..152
3.2.3. Six studies reporting exclusively validation paradigm methods……155
3.2.3.1. Woman Abuse Screening Tool (WAST) – three studies……155

3.2.3.2. WAST-Short and HITS……..157
3.2.3.3. Perinatal Self-Administered Inventory (PSAI)……..158
3.2.3.4. Three question English AAS……..159
3.2.4. Overview of correlation measures……..160
3.2.4.1. Validity evidence based on relations to other variables…….160
3.2.4.2. Validity evidence based on internal structure……162
3.2.5. One study using neither research paradigm……164

3.3. Secondary data analysis results……...165
3.3.1. Commentary……...173

Chapter 4 Discussion……..175
4.1. Overview……175

4.2. What is the evidence for the validity of questions trying to identify IPV in specific ethnic groups?.....176

4.3. Does the evidence for the validity of questions trying to identify IPV vary between different ethnic groups?.....179

4.4. Findings in context of other reviews and clinical practice……182

4.5. Quality Appraisal……186
4.5.1. Evaluating quality appraisal of methodology of the systematic review studies by QUADAS……..186
4.5.2. Outstanding methodological issues related to the systematic review……189
4.5.2.1. Heterogeneity of methods……..189
4.5.2.2. Hierarchy of methods…….191
4.5.3. Evaluating quality appraisal of the use of ethnicity data in the systematic review studies by a five criterion checklist (DECSS)……192
4.5.4. Evaluating the secondary data analysis – methodology and use of ethnicity data…….195

4.6. Limitations of QUADAS……..198
4.6.1. Comparison of QUADAS to STARD……201

4.7. Strengths and limitations of my thesis……204
4.7.1. Strengths…….204
4.7.2. Limitations……..207

Chapter 5 Conclusions and Future Research…….209
5.1. Overview……....209

5.2. Conclusions……...210

5.3. Future research.....211

Figures

Figure 1: The Ecological Model……...21
Figure 2: Health outcomes of intimate partner violence against women…….....29
Figure 3: Categories and subcategories of validity……..50
Figure 4(a): Classical design of a diagnostic accuracy study……...56
Figure 4(b): Results of an accuracy study in the case of a dichotomous index test result…….56
Figure 5: An ROC curve for creatinine kinase values in myocardial infarction……..58
Figure 6: Flow diagram of recruitment of participants to the HARK study……..102
Figure 7: Flow diagram of studies retrieved for systematic review……...109
Figure 8: Receiver operator characteristic curve for the African-Caribbean groups, showing sensitivity of different HARK scores verses 1 – specificity……169
Figure 9: Receiver operator characteristic curve for the south Asian groups, showing sensitivity of different HARK scores verses 1 – specificity……...170
Figure 10: Receiver operator characteristic curve for the white groups, showing sensitivity of different HARK scores verses 1 – specificity……...171

Figure 11: Comparing the three receiver operator characteristic curves in the African-Caribbean, south Asian and white groups……..172

Tables

Table 1: Summary results of 20 studies in systematic review……...110
Table 2: Summary characteristics of 20 studies in systematic review……...115
Table 3: Correlation measures and their interpretation (for criterion correlation validity, convergent validity and association between index scores & external variables)……124
Table 4: Internal consistency reliability measures and their interpretation……126
Table 5: QUADAS quality items……...127
Table 6: Ethnicity quality criteria……...131
Table 7: The sensitivity, specificity, PPV, NPV, LR (with 95% confidence intervals), post-test odds and pre- to post-test probability of IPV at different HARK cut off scores, in the African-Caribbean groups……166
Table 8: The sensitivity, specificity, PPV, NPV, LR (with 95% confidence intervals), post-test odds and pre- to post-test probability of IPV at different HARK cut off scores, in the south Asian groups……...167
Table 9: The sensitivity, specificity, PPV, NPV, LR (with 95% confidence intervals), post-test odds and pre- to post-test probability of IPV at different HARK cut off scores, in the white groups……168

Boxes

Box 1: Five criterion checklist (DECSS) for quality appraisal of the use of ethnicity data……...85
Box 2: STaT (slapped, threatened or thrown) questions……135
Box 3: Partner violence screen questions……...…136
Box 4: Woman Abuse Screening Tool (WAST) questions……138
Box 5: HARK (Humiliation, Afraid, Rape, Kick) questions……139
Box 6: Three Chinese AAS questions……142
Box 7: Portuguese AAS's anchor question on physical IPV during pregnancy……143

Box 8: Ongoing Violence Assessment Tool (OVAT) questions……...144
Box 9: HITS (Hurt, Insult, Threaten, Scream) questions……...145
Box 10: Five non-graphic domestic violence questions…….147
Box 11: Behavioural Risk Factor Surveillance Survey (BRFSS) questions……..149
Box 12: Women's Experience with Battering Scale (WEB) questions…….150
Box 13: Woman Abuse Screening Tool-Short (WAST-short) questions……..155
Box 14: Two Perinatal Self-Administered Inventory questions……158
Box 15: Three questions from the AAS…….160
Box 16: One question from hospital admission protocol……...164

References……...217

Appendices

Appendix A: Reference standards to identify IPV……232
Appendix B: The HARK paper……..235
Appendix C: Search strings……244
Appendix D: Data collection form……258
Appendix E: Secondary data analyses……264

Acknowledgements

First and foremost I would like to thank my 21 month old daughter, Kiran. Her smiling, laughing, bright face kept me working late at night. I am sure that I would never have completed my thesis without Kiran's amazing vision. Pre-Kiran, I planned to leave the UK for adventures overseas which did not include writing up a thesis. So thank you Kiran – I hope that one day you and every little girl in the world can grow up without any fear of violence throughout your lives.

Kiran's unexpected arrival resulted in a change of my timetable – including maternity leave from May 2009 to May 2010 and then working part-time from January 2011. Consequently, there has been a delay in submitting my thesis.

Secondly I would like to thank my Father, Harbhajan. His steadfast belief in me, never doubting my work and his full unwavering emotional support was important. His story of being a refugee boy who crossed the new border alone from Pakistan into India in 1947 is a part of my ethnic identity. My Mother, Harbans, my sister, Kuldip and my Father also provided practical support – making Kiran laugh when I went to college and libraries etc. My other siblings have shaped who I am – their experiences have influenced me.

Professor Gene Feder and Professor Sandra Eldridge were great supervisors, providing valuable input and detailed feedback at every stage of this research. Professor Gene Feder's suggestion that I commence an MD first led me on to this winding path. I am very privileged to have him as my mentor. I hope our job sharing of the Royal College of General Practitioners' Clinical Champion for Domestic Violence role, helps to bring the issue of violence to the forefront of general practice.

I would also like to acknowledge a number of people who helped me with practical tasks throughout my MD research. Alain Besson taught me about my reference manager, dealing with my many queries. David Lexton resolved formatting errors in my thesis. Robert Froud and Stephen Bremner untangled some of the mysteries of STATA (statistical software package), helping me to explore my data. Jan Whalley

improved the presentation of my diagrams. Mark Rose was very generous with his time, sharing his experience of conducting a systematic review. Jean Ramsey helped me to grapple with search strings. Chris Griffiths never became impatient. David Jenkin (my brother-in-law) came to the rescue with a PDF creator that worked. All the women who participated in the studies that I systematically reviewed and my original HARK study – a special thanks.

Finally, I would like to thank my long suffering husband, Hiran. He has had to bear the practical consequences of the many hours that I have spent on this task whilst knowing that he was the first to commend Gene's suggestion to study for an MD. The stories of women who we have met and continue to meet in our separate clinical practices, in east London and around the world, who experience violence should remind us what my work is about. Hence, I dedicate this thesis to all the women who live with intimate partner violence. The intention is not to shock or offend but I include a quotation from one woman in an attempt to evoke what she has endured:

"Fuck you, you fucking bitch."

Glossary

Abbreviations used in thesis:

AAS: Abuse Assessment Screen

ARI: Abuse Risk Inventory

BMJ: British Medical Journal

BRFSS: Behavioural Risk Factor Surveillance Survey

CAGE: Cut down, Annoy, Guilt, Eye-opener questions

CAS: Composite Abuse Scale

CONSORT Consolidated Standards of Reporting Trials

CTS (CTS2): Conflict Tactics Scale (Conflict Tactics Scale Revised)

DECSS: Described, ethnicity, classification, self-assigned, socio-economic questions

FN: False negative result

FP: False positive result

HARK: Humiliation, Afraid, Rape and Kick questions

HITS: Hurt, Insult, Threaten, Scream questions

IPV: Intimate partner violence

ISA (ISA-P): Index of Spouse Abuse (Index of Spouse Abuse Physical dimension)

JAMA: The Journal of the American Medical Association

LR: Likelihood ratio

NPV: Negative predictive value

OVAT: Ongoing Violence Assessment Tool

PCD: Patient Career Diary

PHQ: Patient Health Questionnaire

PPV: Positive predictive value

PSAI: Perinatal Self-Administered Inventory

PTO: Post test odds

PTSD: Post traumatic stress disorder

PVS: Partner Violence Screen

QUADAS: Quality Assessment of Diagnostic Accuracy Studies

ROC: Receiver operator characteristic

STARD*: Sta*ndards for *R*eporting of *D*iagnostic Accuracy statement

STaT: Slapped, Threatened or Thrown questions

TN: True negative result

TP: True positive result

UK: United Kingdom

US: United States

WAST: Woman Abuse Screening Tool

WEB: Women's Experience with Battering Scale

WHO: World Health Organisation

WOMB: Women's views of birth antenatal satisfaction questionnaire

Outputs related to thesis

Publications and papers in preparation

Sohal H, Eldridge S, Feder G. The sensitivity and specificity of four questions (HARK) to identify intimate partner violence: a diagnostic accuracy study in general practice. BMC Fam Pract. 2007; 8:49.

Sohal H, Feder G. 10 minute consultation: Violence between intimate partners. In preparation. For submission to BMJ.

Sohal H, Feder G, Eldridge S. Identifying intimate partner violence in different ethnic groups in primary care. In preparation. For submission to J Interpers Violence.

Sohal H, Eldridge S. Measuring validity: The integration of diagnostic accuracy and traditional validation methods in one framework. In preparation. For submission to BMC Med Res Methodol.

Sohal H, Feder G. DECCS: A five criterion checklist for quality appraisal of the use of ethnicity data in health research. In preparation. For submission to Ethn Health.

Conferences and presentations

Sohal H. What questions identify intimate partner violence in specific ethnic groups in primary care? Presentation to the 2011 Academy on Violence and Abuse Biennial Scientific Assembly: The Developing Science of Violence and Abuse: Toward a New Understanding. Minnesota, April 2011.

Sohal H. My roots: psychometrics and ethnicity in domestic violence research. Seminar presented to Institute of Health Sciences Education. November 2010

Sohal H. Identifying intimate partner violence. Poster presentation at William Harvey Day, Barts and the London School of Medicine and Dentistry. October 2010

Sohal H. Identifying intimate partner violence. Poster presentation at Institute of Health Sciences Education research afternoon. June 2010

Sohal H. The identification of intimate partner violence in primary care, in specific ethnic groups. Seminar presented to Institute of Health Sciences Education. April 2008.

Chapter 1: Background

1.1. Introduction and Chapter Overview

This thesis is concerned with identifying intimate partner violence (IPV) against women in different ethnic groups in primary care. IPV against women is the violence, that is perpetrated by a husband or other intimate male partner against a woman, often termed domestic violence. IPV includes physical, emotional and sexual abuse. For a more comprehensive definition see section 1.3.6.6. My work does not include IPV against men as this is quantitatively and qualitatively different from IPV against women. Repeated coercive, severe physical and / or sexual violence is commoner in IPV against women.[1]

The Ecological Model of IPV (see figure 1, on page 21) lists the factors that influence the use of violence in a relationship. This holistic framework proposes that IPV is the result of individual, relationship, community and societal features that dynamically interact. In the model IPV results when multiple factors from these various spheres exist together and not when only one factor exists from a single sphere.[1, 2] Identifying IPV in primary care is part of a health service response which should be embedded in a wider community response that aims to reduce the level of IPV and its health consequences.

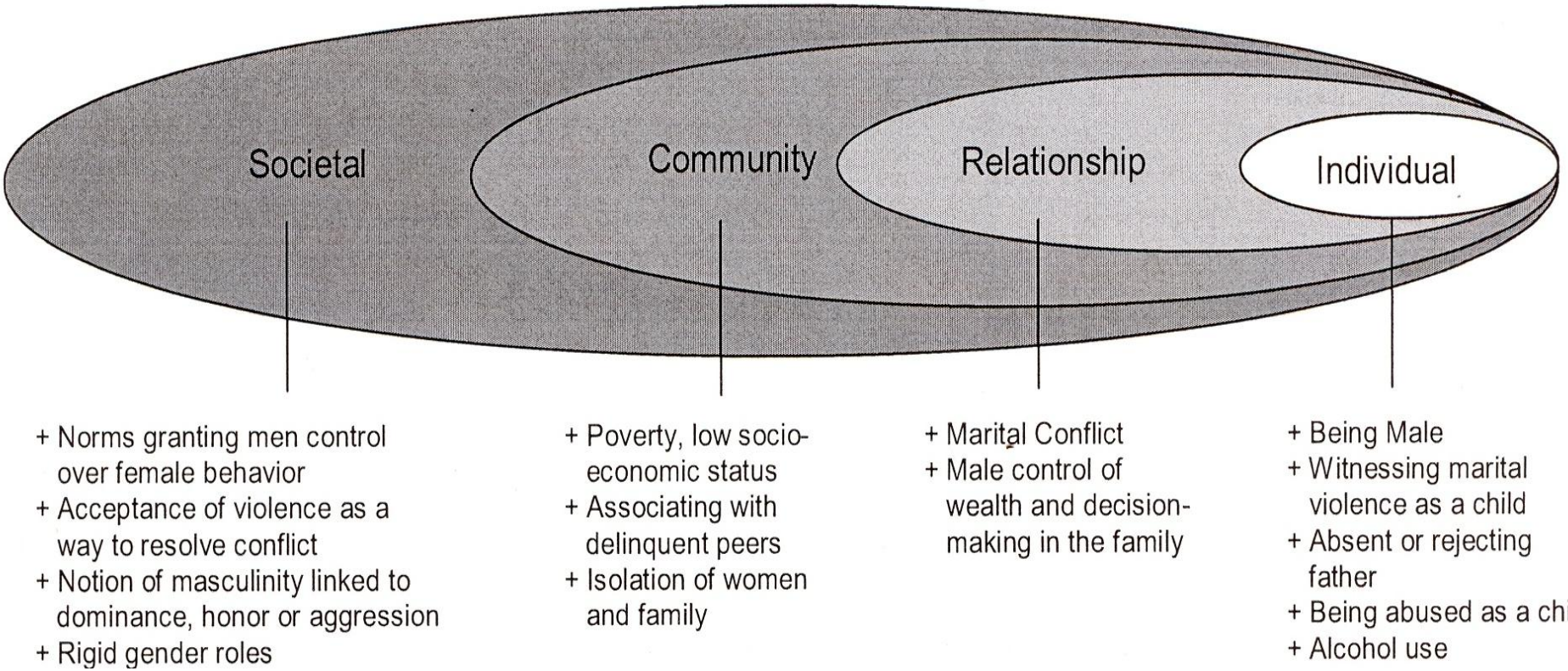


Figure 1: The Ecological Model[2]

This background chapter will describe the theoretical concepts and the literature which underpin this thesis and my research questions (see section 1.5.1.). It will explain the terms used in my research questions, including their intricacies and implications and how and why my two research questions were generated.

In section 1.2., I will examine the case for IPV to be addressed and prioritised by primary care. This is achieved by examining IPV prevalence, the health impact of IPV and the effectiveness of intervening; as well as considering the role of simple valid questions to identify IPV in consultations – particularly during selective clinical enquiry. I will explore whether IPV needs identification in order to potentially prevent rather than just manage the adverse health consequences of IPV. Additionally I will consider how identifying IPV potentially improves the diagnosis of other conditions in primary care.

Following this account of my rationale for this thesis, in the next two sections I will explore the background to measuring validity and understanding the term ethnicity. I will highlight general principles by using examples from IPV research literature. I will show how validity is measured and how the term ethnicity is used in IPV research papers.

In section 1.3., my thesis focuses on the measurement of validity. I will review the different types of evidence that can be used to measure the validity of questions aiming to identify target conditions including IPV. This involves exploring models of validity that originate from distinct disciplines (health sciences and psychometrics) including the classical diagnostic accuracy paradigm, the validation paradigm as well as categorical and dimensional models. I will then integrate these complementary methods into a framework which is based on existing standards. This framework was used to structure the results of my systematic review of research literature describing questions aiming to identify IPV in different ethnic groups. This framework allows an analysis of the diverse evidence for validity presented in these papers.

In section 1.4., I will explore the concept of ethnicity and how it is used in health research studies. I use the phrase “health research” as an umbrella term which

includes epidemiological, clinical and health services research. The theoretical and empirical relationships between ethnicity and race will be examined in this section. I will then expand on the rationale for and potential dangers of collecting ethnicity data in health research studies. Subsequently I will present five criteria to assess the use of ethnicity data by papers.

In section 1.5., I will review IPV research that has used ethnicity data. This leads to the final articulation of my two research questions, followed by an examination of my aims with objectives and an outline of my thesis.

1.2. Intimate partner violence

In this section I review different areas of IPV research in order to examine the case for primary care to address and prioritise IPV. I review studies into IPV prevalence, IPV's impact on health, interventions used when IPV is identified and the approaches used to identify IPV (screening, routine enquiry and selective enquiry).

1.2.1. Is IPV a priority that should be addressed by primary care?

A target condition is an identifiable condition which requires some form of action, for example further tests or treatment changes.[3] It is not necessarily a disease. Intimate partner violence (IPV) could be considered a target condition which on identifying should prompt further action by the health care professional (see section 1.2.1.4.2.2, page 37). However to consider IPV as a priority health target condition that should be addressed by primary care, means that IPV needs to be identified even when a woman has no direct physical injuries caused by IPV or accompanying illnesses. Therefore health care professionals would need to directly consider and manage IPV itself. This is distinct from treating injuries caused by IPV and / or managing the health consequences of IPV, (for example depression).

Firstly I consider the prevalence of IPV by introducing a large global World Health Organisation (WHO) study and examining a systematic review from the United Kingdom (UK) which demonstrated prevalence differences between community and clinical populations.

1.2.1.1. Prevalence

Violence against women is a global issue affecting millions who experience it and have to live with its consequences.[4] The WHO Violence Against Women study[5] found that the prevalence of lifetime physical violence and sexual violence by an intimate partner, among ever-partnered women, varied from 15 to 71% in urban and rural settings in 10 countries.[6] For more details on this study, see sections 1.2.1.2. and 1.5.

A recent systematic review of prevalence studies in the UK found that in community surveys lifetime prevalence of IPV varied from 13 to 31% whilst in clinical populations it ranged from 13 to 35% with the highest levels found in women presenting to Accident and Emergency Departments.[7] IPV prevalence from different studies was difficult to compare due to variations in the study population, study setting, study designs (self completed and researcher completed questionnaires), time frames, age of participants and the definition of IPV used. Some studies included physical, sexual and emotional IPV whilst most frequently studies only focussed on physical IPV.

The 15 UK prevalence studies in this systematic review confirmed that study population (community verses clinical) was associated with a variation in prevalence. Community populations had significantly lower IPV prevalence.[7] This was exemplified by two of the studies from this systematic review. A computerised self completion method in a nationally representative sample of 24,498 women and men showed that the adjusted lifetime prevalence of physical, emotional, financial abuse, threats or force was 25% and the incidence was 5%. The lifetime prevalence of sexual assault was 23% and the incidence was 3%.[8] In contrast, a study by Richardson and colleagues of 1,207 women attending general practice found a physical IPV lifetime prevalence of 41% and IPV incidence of 17%.[9]

The prevalence of IPV tends to be higher in women attending health care services than in those participating in community surveys even when these studies are set in the same geographic population.[10] These prevalence studies clearly indicate that

IPV is common in clinical populations. This includes primary care but also in other clinical settings. For example, the lifetime prevalence of severe domestic violence experienced by psychiatric inpatients is between 30 to 60%.[11] I now examine studies that have looked at the health impact of IPV.

1.2.1.2. Health Impact

The National Center for Injury Prevention and Control in the US reports that 5.3 million episodes of domestic violence occur each year, causing 2 million injuries with 550,000 requiring medical treatment.[12] In the UK, two women are killed by their current or former partner each week.[13] Apart from the obvious immediate effects of physical injuries, IPV also causes other short and long term health problems.

The WHO Violence Against Women study comprehensively measured the health impact of IPV around the world.[14] Following interviews with 24,097 women in ten countries, pooled analysis of all sites found significant associations between lifetime IPV experiences and suicidal attempts (3.8 [95% CI 3.3-4.5]), suicidal thoughts (2.9 [95% CI 2.7-3.2]), vaginal discharge (1.8 [95% CI 1.7-2.0]), memory loss (1.8 [95% CI 1.6-2.0]), dizziness (1.7 [95% CI 1.6-1.8]), pain (1.6 [95% CI 1.5-1.7]) and difficulty with daily activities (1.6 [95% CI 1.5-1.8]). Other controlled studies from a wide range of settings, have also shown associations with gynaecological conditions (including sexually transmitted diseases) and chronic pain as well as gastrointestinal conditions.[15]

A cross-sectional study has shown that IPV was associated with 8% of the overall disease burden in women aged between 18 to 44 years in Victoria, Australia. 73% of the disease burden attributed to IPV was due to poor mental health (depression, anxiety and suicide) and 22% due to substance abuse (tobacco, alcohol and illicit drug use).[16] In women aged less than 45 years, IPV was the most important risk factor out of the eight major risk factors for ill health. These risk factors included high blood pressure, high cholesterol and body weight. IPV was double the risk of illicit drug

use, the risk factor closest to it which contributed less than 4% of the disease burden. High blood pressure only accounted for 1% of the disease burden in this age group.

Many of the health impact studies of IPV focus on the effects of physical IPV. However psychological IPV has also been found to produce long term adverse physical and mental health effects.[17] Coker and colleagues found using a random digit dial telephone survey of 13,912 women and men aged between 18 to 65 years that logistic regression models that included both psychological and physical IPV scores, higher psychological IPV scores were more strongly associated with current poor health, depressive symptoms, substance abuse and developing either a chronic disease, chronic mental illness or an injury.[17]

More recently Yoshihama and colleagues,[18] engaging with the WHO's cross-national research endeavour, in Japan found that the impact of emotional IPV only was similar to the impact of emotional IPV with physical or sexual IPV. They concluded that health care professionals needed greater awareness about the effects of emotional IPV. Ludemir and colleagues,[19] found that during pregnancy psychological IPV was strongly associated with postnatal depression which was independent of both physical and sexual IPV. Sexual IPV has also been shown to be a separate dimension of IPV which can occur with or without physical IPV.[20, 21]

A systematic review of IPV health impact studies showed that IPV significantly increased the risk of mental illness and substance abuse.[7] This systematic review included a meta-analysis which examined actual physical IPV and threats of physical force as risk factors for mental health problems in women.[22] In this meta-analysis, the strength of association as well as temporality was examined. A significant association was found between physical IPV and depression, post-traumatic stress disorder (PTSD), suicide, suicidal thoughts, alcohol abuse and drug abuse. Depression and PTSD were the most frequent mental health sequelae of physical IPV. When physical IPV stopped, depression decreased. Both depression and PTSD reacted to whether physical IPV was present or absent. Additionally a dose-response relationship showed that physical IPV's severity and duration was associated with depression and PTSD's severity and prevalence. This suggests a causal relationship

between physical IPV and adverse mental health outcomes.[22] In women attending general practice physical, emotional and sexual abuse has been found to be associated with depression.[23] More recently, it was found in general practice that women who had ever been afraid of a partner on average had higher depressive symptom scores than women who had never been afraid.[24] Additionally, increased psychotic symptoms have been shown to be related to assault including domestic violence.[25] The full array of potential health outcomes of IPV against women are listed in figure 2, on page 29.

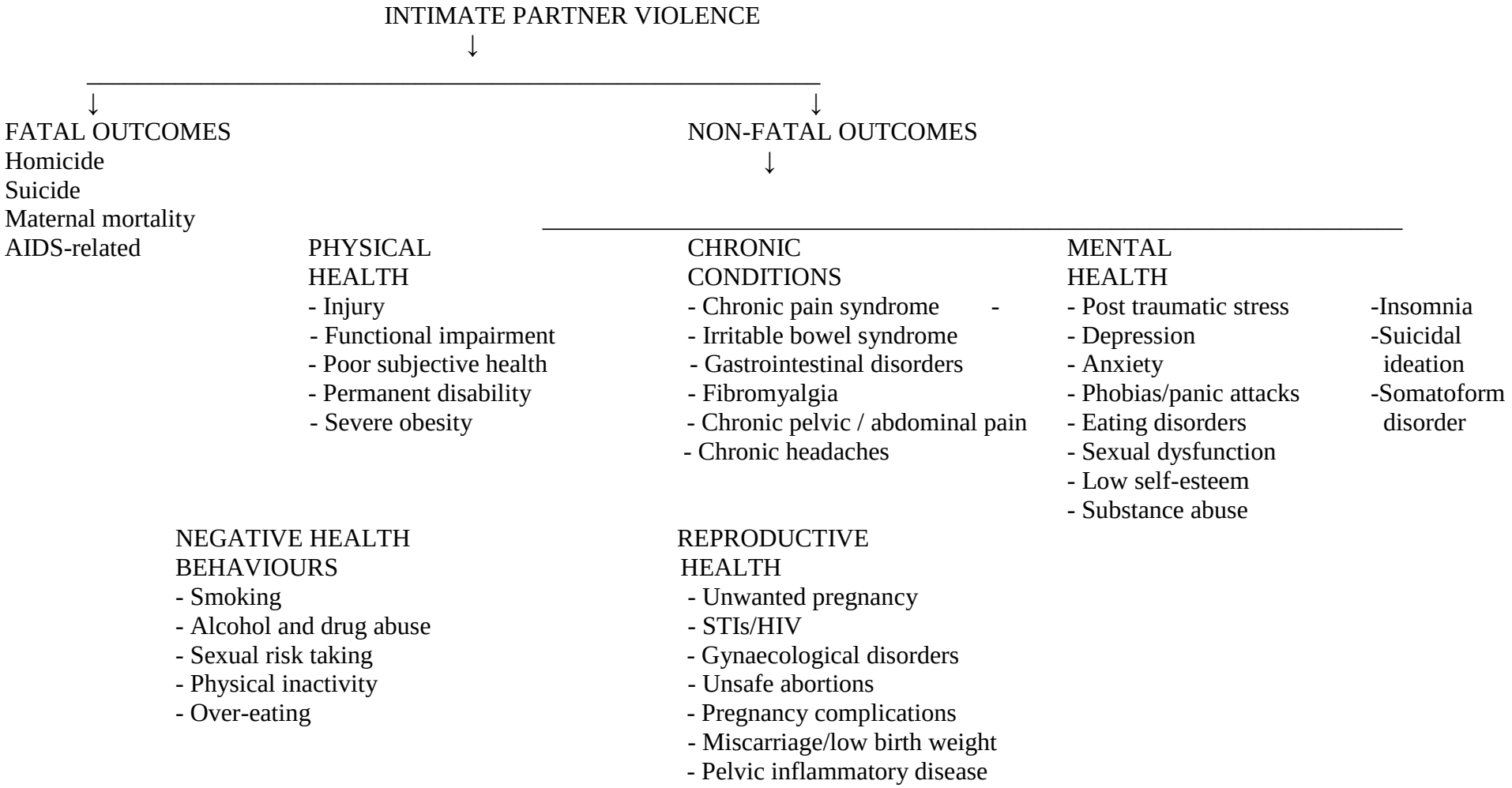


Figure 2: Health outcomes of intimate partner violence against women[1, 15, 128]

There have been numerous studies looking at the health impact of IPV on the children within affected families. A recent longitudinal cohort study showed that children whose mothers had experienced IPV had higher health care utilization and costs, even if their mothers' abuse stopped before they were born.[26] IPV has been found to be an independent risk factor for deficit in gestational weight gain during pregnancy[27] with evidence of IPV being associated with low birth weight.[28] In the developing world it has been suggested that IPV is a factor in under two year old mortality as well as child malnutrition.[29] Children exposed to severe and recurrent IPV are more likely to be admitted with acute malnutrition.[30] Many children live with IPV with negative impacts on their health and development.[31] Children exposed to IPV often experience emotional and behavioural problems.[32] Identifying IPV within families containing children requires child protection issues to be examined.

A case has been made for health care professionals identifying, prioritising and managing women experiencing IPV on the basis of the evidence described so far. This shows that IPV is a major public health problem, as it is common and associated with multiple health conditions including a detrimental effect on children's health. Additionally, as women affected by IPV are frequently isolated, health care professionals are in a unique position often being the only professional to have contact with these women.[33, 34]

However it has also been argued that high IPV prevalence, IPV's considerable health impact and health care professionals' distinct role in potentially identifying IPV are not sufficient to transform IPV into a condition that should be addressed directly or prioritised by primary care. Hence IPV researchers have looked for further evidence of benefit for when health care professionals identify and intervene in IPV, contending that this is required if IPV itself, is to be recognised as a priority health target condition with

"a specific role for health services in responding to it other than in the management of its health consequences."[33, 35]

This further evidence includes research into the interventions that can be offered when women are experiencing IPV. This has looked at whether the adverse health outcomes

which can accompany IPV can be prevented. I now summarise the body of research into IPV interventions.

1.2.1.3. IPV Interventions

It has been proposed that if IPV can be identified early, interventions could be developed to prevent adverse mental and physical health conditions.[17] Evidence from a systematic review originally suggested that in women experiencing current IPV either who have actively sought help or are in a refuge, referral to a domestic violence advocate may decrease abuse, increase quality of life, social support and safety behaviours. [36] The most recent Cochrane systematic review based only on randomised controlled trials comparing advocacy interventions for women with a history of IPV against usual care found that intensive advocacy (for 12 hours or more) could reduce physical IPV one to two years after the advocacy intervention for women recruited in domestic violence shelters or refuges.[37] It is unknown whether intensive advocacy has a beneficial effect on these women's quality of life or mental health; or whether less intensive advocacy in healthcare settings for women living with the perpetrators of violence is effective. Psychological interventions may be effective for women who have also either sought help or been in a refuge as opposed to those who disclose on screening.[7] System based interventions involving staff training, clinician prompts, referral measures, waiting room posters and audit found increased identification of IPV and referral to domestic violence services.[36] There is little evidence for the effectiveness of giving advice on safety planning and behaviour.[38, 39] Parenting interventions with female survivors and their children improved behavioural and emotional outcomes for both mothers and their children.[40]

Overall, research into interventions does seem to suggest that some of the potential gain from preventing IPV's adverse health outcomes is achievable. However this is based on evidence for women who are actively seeking help as there is currently an absence of evidence for women who have been identified pro-actively by health care professionals. A study that does identify abused women who are pregnant or with

infants, in primary care will be reporting on the clinical outcomes of general health and depression following an IPV intervention of 12 months of non-professional but trained and supported mentor mother support.[41] I will now examine the research into the different approaches used to identify IPV.

1.2.1.4. Approaches to identifying IPV

In this section I will review the approaches used to identify IPV (firstly screening, then routine enquiry and lastly selective enquiry) in order to continue assessing the case for primary care to address and prioritise IPV. The relationship (i.e. the supposed dichotomy) between screening and routine enquiry will be considered (see section 1.2.1.4.2.1.).

1.2.1.4.1. Screening

Effective IPV screening can be defined as a process whereby those who don't necessarily perceive that they are at risk of IPV as well as those already affected by IPV or its sequelae, are asked a question, to identify individuals who are more likely to be helped than harmed by an intervention to reduce the risk of IPV.[42] IPV screening can only be promoted if it improves health outcomes for women. The Wilson screening criteria are a list of principles that should ideally be fulfilled by an effective screening programme,[43] including an IPV screening programme. There has been wide debate and research examining whether screening for IPV is beneficial.[44] This has partly been driven by the many US professional organisations which advocated the introduction of IPV universal screening without evidence to support this recommendation.[45-49] Overall the conclusion from systematic reviews is that there is currently inadequate evidence to support universal screening in health settings.[50-52]

1.2.1.4.2. Enquiry

Compared to IPV prevalence, IPV's health impact, IPV interventions and IPV screening, there has been much less quantitative research looking at the effects of enquiring (both routinely and selectively) about IPV.

1.2.1.4.2.1. Routine enquiry

Routine enquiry has been advocated by those who have rejected the public health approach of IPV screening.[53] They have argued that it is inappropriate to apply the Wilson screening criteria when IPV is not a medical illness that needs to be diagnosed but rather a health-related risk factor that needs to be identified, in the same way that smoking, and obesity are identified in general practice, i.e. by using regular and repeated enquiries in routine consultations.[33, 54] Proponents of routine enquiry contend that it has a broader remit than screening in that identification via disclosure is not the principal aim with less pressure for women to disclose IPV. Instead routine enquiry's purported added benefits are that it can be used as a vehicle to provide supportive information about IPV to women whilst simultaneously decreasing stigma[55] and changing society's attitudes towards IPV.

I think that these goals would apply equally to a well constructed screening program. In effect renaming screening, routine enquiry, does not resolve that the Wilson screening criteria are not supported by current existing evidence.[56].There is still a need to show that both screening and / or routine enquiry are effective and safe.[35] There is sometimes an assumption that as asking questions is not an invasive test, that it must be a safe exercise. This was not borne out in a qualitative study which showed that some general practitioners managed IPV disclosure in a potentially unsafe way – breaking confidentiality and undertaking couple counselling.[57]

1.2.1.4.2.2. Selective enquiry

Selective enquiry has also been referred to in IPV literature as targeted identification,[35] trigger enquiry, case finding[58] and diagnostic evaluation.[59] Asking about IPV and identifying IPV facilitates the diagnostic process. This is

different from the routine enquiry of all women and IPV screening, as well as being quite separate from offering IPV interventions. Despite the lack of quantitative evidence, identifying IPV independent of interventions has been recognised as an important task [52, 60] which is a part of good quality clinical care. I now expand on the benefits of selective enquiry.

The process of hypothetico-deductive reasoning[61] that is undertaken in arriving at a likely diagnosis for the cause of the patient's symptoms is affected by knowing that a woman is experiencing IPV. For example, in order to decide whether a woman's symptom of chronic pelvic pain with deep dyspareunia is more likely to represent endometriosis, (a pathological diagnosis signifying the presence of endometrial glands and stroma outside of the endometrial cavity) or to be related to IPV requires identifying whether IPV is present. Chronic pelvic pain with deep dyspareunia could be related to IPV either indirectly due to emotional distress or possibly more directly due to soft tissue inflammation from repeated forced sex. Regardless of whether coercion is involved the woman may not enjoy having sex with her partner who is abusive in other ways which has led to a medicalisation of her symptoms. Identifying IPV may avoid the cycle of repeated gynaecological referrals, invasive tests for example, laparoscopy with or without biopsies whilst the gynaecologist has to try to differentiate between symptomatic and asymptomatic endometriosis. Attributing a woman's symptoms to the diagnosis of endometriosis is aided by knowing whether IPV exists. A woman experiencing IPV may indeed have symptomatic endometriosis which requires treatment but the clinical picture cannot be fully judged without a complete history, including the social history and a frank discussion with the woman about her personal circumstances. In at least one third of women with chronic pelvic pain, no organic cause is found on laparoscopy.[62] Most recently in 487 women with chronic pelvic pain, 70% had no endometriosis on diagnostic laparoscopy with 55% having no obvious pathology.[63] Instead psychosocial factors, including abuse, have been found to be strongly associated with chronic pelvic pain, including dyspareunia.[64]

The psycho-social context always affects symptoms and how they are expressed, hence the importance of the social history when taking a medical history. An

important part of the social history is whether a woman is experiencing IPV. The association between social circumstances and disease has long been recognised.[65] Most recently a WHO report on health inequity recognized the unequal distribution of power, including gender inequity underlying poor health.[66] At the grassroots level mental health professionals have queried whether asking about IPV is relevant to the history taken when assessing mental health.[67]

I would argue that the clinician having knowledge of whether IPV exists, following selective enquiry as part of a diagnostic assessment, directly affects the diagnosis or exclusion of some conditions (for example anxiety) as well as management of conditions (for example depression). Hence, a woman who presents with mild or moderate depression and is experiencing IPV needs to have this identified. This can then allow a detailed examination of the woman's situation and her own resources. Prescribing antidepressants may serve to only circumvent this pertinent discussion. The importance of careful interpretation of research findings to guide treatment in individual cases of depression seen in clinical practice has been highlighted.[68] Identifying exposure to IPV is also central in potential cases of child protection.[69]

The majority of women experiencing IPV do not present to primary care with acute injuries. Instead they are far more likely to present with medically unexplained, non-specific symptoms such as chronic pain (headache, abdominal pain and gynaecological pain), mood disturbances (anxiety, depression) or addiction (alcohol and other drugs).[70, 71] Women may choose not to disclose their experience of IPV for a number of reasons. They may find it difficult to disclose unless they are specifically asked by health care professionals. They may decide that non-disclosure is appropriate after assessing the risks and benefits to them personally of disclosure.[67, 72] Or they may think that it is not relevant to disclose IPV. This may occur if women do not make a connection between their symptoms and their experience of IPV.

Health care professionals as well as women attending primary care need to be able to make a link between medically unexplained symptoms (for example headaches, abdominal pain, dyspareunia) and IPV, as well as mental health conditions (for

example depression and anxiety). The bio-psycho-socio-immunological mechanisms that may be at play in IPV have been described.[73] Less well known than the direct effects of trauma mediated via physical and / or sexual IPV, are the indirect stress effects which are thought to be mediated via the over-responsiveness of the autonomic nervous system, with the sympathetic nervous system producing excess, un-modulated stress hormones. Tension headaches may result due to increased muscle contraction caused by the sympathetic nervous system. Migraine due to vasoconstriction mediated by increased norepinephrine and serotonin levels, followed by rapid vasodilatation and pain. Sustained hypertension could be related to increased peripheral vascular resistance mediated via increased alpha-adrenergic tone during chronic stress. Cortisol, catecholamines, cytokines and Th cell balance are thought to be related to depression and post traumatic stress disorder (PTSD). It has been shown in a small group of women who did not smoke, abuse drugs or alcohol, were not pregnant or medically ill but did have PTSD with a history of childhood sexual trauma, that they had significant increased immune activation, demonstrated by increased biological markers (CD45RO / CD45RA lymphocyte ratio) compared to matched controls.[74] In women with lifetime IPV related PTSD, salivary cortisol was raised compared to women exposed to IPV who did not develop PTSD.[75] In the offspring of rodents who were repeatedly stressed during gestation, structural alterations have been shown in their brains and their hypothalamic pituitary adrenal axis.[76] The importance of further research into how the experience of IPV changes psychological, biological, neurological, behavioural and physiological pathways has been recognised.[77]

Identifying that IPV is occurring may be a fundamental step enabling health care professionals and their patients to understand previously unexplained presentations.[78] A health care professional suggesting a link between a woman's symptoms and her experience of IPV may help patients to begin to deal with their predicament rather than avoiding it by proceeding into a cycle of repeated referrals and investigations. For example, a patient with chronic unexplained headache should at some point trigger an enquiry about IPV with an aim to avoid inappropriate investigations or treatments[59] that fail to address the underlying issue.[79] This approach may avoid inappropriate referrals (for example to a neurologist). Another

example is of a young woman with insomnia. This may also trigger asking questions about IPV, perhaps avoiding ill thought out treatment for sleep disorders with benzodiazepines. Asking about IPV and identifying IPV potentially facilitates the diagnostic process.

Identifying IPV should prompt further action by the health care professional.[72] According to expert consensus opinion,[71] this should ideally include carefully listening to the woman (whilst reassuring her that IPV is not her fault and that it is common), discussing safety planning, considering her children, making her aware of local and national support services (including the domestic violence and refuge 24 hour helpline) and providing follow up. It is important that each woman is made aware of her options whilst at the same time being supported in the decisions that she makes.[53]

A health care professional's support over the long term may help the woman in being able to change her own situation[71, 80] even without formal referrals. Health care professionals who respond appropriately to women who are identified as experiencing IPV (i.e. listening, being non-judgmental, compassionate, caring, and confidential whilst validating the woman's experiences) facilitate this process despite women rejecting intervention or referral to an external agency.

The research into IPV prevalence, IPV health impact, IPV interventions and the important benefits of selective enquiry in identifying IPV, as described above, all support that IPV should be addressed and prioritised by primary care. My review supports identifying IPV for its own sake in order to potentially prevent the adverse health consequences of IPV (which is quite separate from only managing the adverse health consequences of IPV once they are present) and to improve clinical diagnosis generally. Regardless of whether IPV is identified using selective clinical enquiry, routine enquiry or screening, simple, brief and valid questions are required that can be used in consultations to identify women who are or have experienced IPV.

Valid questions that can identify women from different ethnic groups experiencing IPV who present in clinical settings are a pre-requisite for an appropriate response from health services to this substantial public health problem.

I have examined the case for IPV to be prioritised and addressed by primary care whilst highlighting the importance of identifying IPV. I now consider the background to measuring the validity of questions.

1.3. Measuring Validity

Throughout this section, most prominently in my account of the five integrated categories of validity evidence (see section 1.3.6.), I will employ questions used to identify IPV as examples. This aids understanding of how validity can be measured.

In this section, I will first look at the role of measurement in scientific research, focussing on how measurement is used in the health sciences to evaluate medical tests, including questions which are part of a clinical history. I will then contrast how questions have been evaluated in different disciplines particularly the approach taken to measurement error. This leads to a description of the diagnostic accuracy paradigm and the validation paradigm both of which can be used to establish whether questions are valid, i.e. are measuring what they are supposed to measure. Categorical and dimensional models are then considered which capture the differences between the diagnostic accuracy and validation paradigms. Multi-dimensional scaling incorporates categorical and dimensional models. This union is mirrored by the integration of the diagnostic accuracy and validation paradigms within the 1999 Standards for Educational and Psychological Testing. These Standards represent a logical categorisation of the disparate body of evidence which can be used to describe the validity of questions. Five sub-sections then describe the five categories of validity evidence which are described within the Standards. These are based on the consequences of testing, relations to other variables (which include criterion performance studies, i.e. diagnostic accuracy and criterion correlation studies, association studies and known group comparisons), internal structure, response processes and test content. A separate sub-section (1.3.6.4.) draws attention to correlation, a statistical method which features heavily in a number of the methods used to measure validity (criterion correlation studies, association studies and internal consistency reliability). Attention will be drawn to the commonality of different correlation coefficients.

1.3.1. Measurement

Measurement is central to all quantitative scientific research regardless of whether it is in the natural sciences, social sciences or health sciences. Different disciplines have often developed quite separate ways of looking at the common problem of measurement without engaging with methods and theories in other fields.[81]

In health sciences, medical tests are often evaluated to see if they are measuring what they are supposed to measure. This is the classical definition of validity. Medical tests can identify physiological derangements, establish prognosis, monitor illness, diagnose illness or identify target conditions.[82] A medical test does not just have to be a biochemical blood investigation, microbiological test (for example urine culture) or imaging study (for example chest radiograph). A medical test can also refer to questions when taking a patient's history or a manoeuvre when performing a clinical examination. The potential power of the clinical information collected in a consultation to make a diagnosis has been strongly argued by Sackett and colleagues,[61, 83] who highlighted the potential of simple clinical observations not only to inform diagnosis and therapeutic responsiveness but possibly also to ascertain prognosis.

Hence it is important to know whether questions in histories or manoeuvres during examinations measure what they are supposed to be measuring or to know what their measurements mean. This is especially true in resource poor environments where one may not have recourse to further expensive technical investigations or even in well resourced environments where a definitive diagnostic investigation, a so-called gold standard, may not exist. In these two scenarios health care professionals may rely primarily on the clinical history and examination to guide management of the patient.

1.3.1.1. Questions and their evaluation

Questions used for measuring have been used and evaluated differently in the fields of health sciences, psychometrics, psychology (including measurement of intelligence and personality) and education. I will now consider the differences and similarities in the way these questions are used in medicine and psychometrics.

In medicine, questions are central to taking a clinical history and the first part of any clinical encounter with a patient. The history helps to formulate a short list of possible diagnoses, also known as the differential diagnosis. Diagnosis has commonly referred to a disease or illness but diagnosis is now also conceptualised as identifying a target condition. A target condition rather than just meaning a disease can also include any identifiable condition which requires some form of action, for example further tests or treatment changes.[3] There are many aspects of taking a history in which identifying target conditions (for example, whether a patient is a smoker) is separate but can be as important as making a diagnosis. Intimate partner violence (IPV) is not a disease but refers to a social issue which can also be a risk factor for poor health. Therefore IPV can be considered a target condition which on identifying should prompt further action by the health care professional (see section 1.2.1.4.2.2., page 37).

In psychometrics, questions are also used in a variety of ways but not in the context of a clinical history. For example educationalists may use questions in an exam to separate out students with different grades according to their ability in a subject. Market researchers may use questions addressed to the general public to decide the name of a new product. In contrast to a clinical history, participants' responses to often closed questions are utilised in a variety of formats including written questionnaires, computer presentations or face to face interviews. These questions are individually often referred to as items whilst if grouped together to measure one entity they can be called a scale, an assessment tool, a toolkit, an instrument or a questionnaire. This variety of terms all refers to a group of questions.

Similarly in clinical histories, open and closed questions are also often grouped together in different ways for example, the history of the presenting complaint, or the

social history; and in system based groups for example, questions about the respiratory system or more specific sets of questions to make a possible diagnosis such as asthma in children. Questions can be grouped together to identify more precise target conditions for instance asthma related to cat fur in children. Structured parts of histories can also be thought of as a scale or a measurement tool, in the same way that questions are viewed in the discipline of psychometrics.

In psychometrics, the measurement error associated with using questions to measure an entity is formalised with an assumption that any response to a question is subject to an error. Respondents may interpret questions differently which may only partially be accounted for by misinterpreting the question. They may also respond to questions in a biased manner depending on the exact wording of the question or make a mistake writing the answer on to the answer sheet. Within psychometrics there are established methods to try to reduce the measurement error which involves scrutiny of individual items (questions) and the whole scale (group of questions). It has been argued that this provides more valid and reliable information than the information generated by questions in a typical clinical history.[81]

Health care professionals often rely heavily on their personal clinical skills to diagnose conditions, rather than using measures and questions that have been psychometrically tested. This approach has been criticised due to its reliance on the clinical skills of individual health care professionals. For example, if the health care professional has helped the patient to feel comfortable and relaxed, the history obtained will be more reliable. It has been thought that too little consideration has been given to whether the questions used in a history are psychometrically valid.[81] Clinical disagreements have been demonstrated in histories. For example, when 57 men complaining of chest pain were interviewed by three cardiologists, it was found that if one cardiologist diagnosed that a patient had angina, the other two only agreed with him 55% of the time.[84] The need for reproducible, reliable and accurate clinical measurement has been recognised by biomedical researchers with an emphasis on finding out what data is relevant and worth seeking out in the history as well as what is best to ignore.[61] In general practice, the introduction of the Patient Health Questionnaire (PHQ) represents an attempt to provide valid and reliable

questions with which to assess severity and monitor clinical depression, as opposed to just using a clinical history and clinical judgement.[85-87] Psychiatry has also made use of psychometric principles of validity in structured interviews, for example the Diagnostic Interview Schedule.[88] Other health researchers trying to measure what was thought to be un-measurable including subjective states and quality of life in a valid and reliable manner have also turned to psychometrics.[89]

In both medicine and psychometrics establishing that a question is identifying what it is supposed to identify needs evidence. This can be crystallised down into two basic types of research methodologies encapsulated by the diagnostic accuracy paradigm and the validation paradigm. Both paradigms are trying to measure validity. I now describe these two paradigms in more detail below.

1.3.2. Classical diagnostic accuracy paradigm

The phrase "classical diagnostic accuracy paradigm" was recently used in a methodological review study.[3] "Paradigm" is defined by the Oxford Dictionary as "an example or pattern, especially one underlying a theory or methodology."[90] In the present context, the use of "paradigm" emphasises that all diagnostic accuracy studies depend on preceding empirical research providing evidence to develop and support the use of the reference standard. The reference standard is equivalent to the criterion used in criterion validity (see section 1.3.6.2.1.).

Diagnostic accuracy involves comparing the results of the test under evaluation (an index test which I will refer to as the index questions) to a reference standard. Diagnostic indices are then generated (such as sensitivity and specificity) which express how well the index questions are able to identify those with the target condition as classified by the reference standard.[91]

Yerushalmy published the first paper assessing the performance of a medical test using sensitivity and specificity whilst referring to accuracy in 1947.[92] Over the years medical researchers have predominantly depended on this paradigm[93] with

little change to it apart from more accuracy indices being devised including predictive values, likelihood ratios and diagnostic odds ratios.[94]

Central to diagnostic accuracy studies is the role of the reference standard which has to decide whether the target condition is present or absent in all participants. The reference standard is a test which is either known to be able to determine whether a target condition is present or absent without errors (i.e. a gold standard) or more pragmatically, it is the best existing method at determining whether the target condition is present or absent. Hence diagnostic accuracy studies invariably depend on preceding empirical research that has provided evidence to support the use of the reference standard.

The philosophy of applying a diagnostic accuracy model to IPV, a social issue which cannot be defined by a perfect gold standard can be debated. However IPV is like many if not all medical conditions for which a perfect reference standard (i.e. a gold standard) does not exist. Instead satisfactory reference standards have been developed for IPV identification. These are invariably a long set of questions, normally used in and devised for research settings. Researchers often try to improve existing unsatisfactory reference standards prior to embarking on diagnostic accuracy studies. This evolution of reference standards can be seen in the IPV field.[95]

The majority of diagnostic accuracy studies have focussed on investigations as opposed to evaluating questions that form part of a clinical history. There are exceptions. For example, it has been shown that when answering yes to three or more of the CAGE questions (cut down, annoy, guilt, eye-opener), there is a likelihood ratio of 250 for alcohol dependency or abuse.[96] The CAGE questions have been found to be more predictive of alcohol dependence than computer-assisted laboratory data profiles.[97] Smoking for more than 40 pack-years (likelihood ratio 8.3), having a self-reported history of chronic obstructive airways disease (likelihood ratio 7.3) and age over 44 years (likelihood ratio 1.3) are significantly associated with the diagnosis of obstructive airways disease.[98]

1.3.3. Validation paradigm

The term "validation paradigm" in relation to evaluating diagnostic tests was also used in a recent methodological review study.[3] The underlying characteristic of this methodology is that there is no existing high quality reference standard. This results in the central measurement challenge. Therefore unlike diagnostic accuracy studies, studies using the validation paradigm are not necessarily based on earlier empirical research into a reference standard.

Test validation as a concept has been known and used for years. Validity has been traditionally divided into content, construct and criterion validity which were seen as relatively independent characteristics of a measure that needed to be autonomously determined. This use of terminology has evolved (see section 1.3.6.).

Conventional methods originating from the arena of psychometrics, psychology and social sciences have been commonly used to evaluate questions which endeavour to measure or tap into latent traits, for example depression or anxiety. There is extensive theory about the use of questionnaires and their validity.[81]

The Women's views of birth (WOMB) antenatal satisfaction questionnaire is an example of a health tool, developed using psychometric methods from the validation paradigm including examining traditional face, content and construct validity as well as internal consistency reliability.[99] Baker developed the Patient Career Diary (PCD), a measure of patients' attitudes towards health care (at the interface between primary and secondary care) also using face validity, construct validity and internal consistency reliability.[100] Neither the WOMB nor the PCD contained questions that were designed to be used within a clinical history in order to diagnose an illness or identify a target condition.

When evaluating diagnostic tests (including questions) with no acceptable reference standard, applying the concept of a clinical test validation could provide a significant methodological advantage over the traditional diagnostic test accuracy paradigm.[3] The validation process uses a variety of methods to try to establish whether questions

can serve their purpose by exploring meaningful relations between index test results and other relevant clinical characteristics.

The current reference standards that identify IPV are long questionnaires (see Appendix A). These have been developed using a variety of methods which are a part of the validation paradigm, for example factor analysis used in the development of the Composite Abuse Screen (CAS).[101] This use of the validation paradigm has occurred as there can never be an absolute gold standard for identifying IPV. This is because IPV is an opaque entity which probably means many different things to individuals. Therefore though a question, for example asking about abuse, may purport to measure IPV it could actually be irrelevant in identifying IPV (see section 1.5.1.). Additionally, identifying IPV depends on a woman's willingness to disclose her experience of IPV to a health care professional. This though is in common with all questions in a history (for example when obtaining a sexual history).

1.3.4. Categorical and dimensional models

An alternative way of conceptualising the differences between the diagnostic accuracy paradigm and the validation paradigm is by considering the categorical and dimensional models which were summarised by Devins.[102] The categorical model in common with the diagnostic accuracy paradigm has a clear division between cases and non-cases. In the dimensional model, "caseness" is a matter of degree with no clear separating boundary between cases and non-cases. It is the theoretical basis of a construct which should determine whether a categorical or dimensional model is the more appropriate representation of any particular construct.[81] By construct I mean a hypothetical unifying variable. A construct has been thought of as a "mini-theory" in order to explain the relationships among various behaviours (or attitudes).[103] For example, IPV can be interpreted as a construct that helps to explain the connections between physical, sexual and emotional violence. A construct may underlie a cluster of related questions.[89] Construct is often used interchangeably with the terms "dimension," "domain," "area," "attribute," "trait" and "concept."

For a construct which varies quantitatively and qualitatively at different severities it is most apt to use the categorical model (for example urinary tract infection diagnosis based on culture). A categorical construct's severity would be lowest in instances that minimally satisfy diagnostic criteria whereas individuals labelled as non-cases would be free of the construct and in effect not have the disorder. For a construct which varies only quantitatively at different severities it is most suitable to use the dimensional model. Sometimes the understanding of a construct may change so that the most appropriate model used to conceptualise the construct alters. For example, hypertension was once treated as a categorical construct in that a diastolic blood pressure of less than 90mm Hg was deemed normotensive, not requiring treatment whilst a diastolic blood pressure of more than or equal to 90mm Hg was hypertensive, requiring treatment. Now an enhanced understanding of hypertension and how this impacts on health outcomes has resulted in an individual's cardiovascular risk affecting the level of blood pressure at which treatment is initiated.[104] Therefore different actions are required at different blood pressure levels. Hence blood pressure is now treated more as a dimensional construct, with a continuum, as opposed to only having a categorical structure, dividing the population into those with hypertension and those without.[81]

Multidimensional scaling represents an endeavour to bridge these 2 models. It permits a variety of attributes to be measured dimensionally, in such a way that results can be used to both categorise and determine the extent to which these categories are present.[81] The development of the Composite Abuse Scale, a reference standard used to identify IPV, made use of multidimensional scaling.[101] The four dimensions of IPV defined by the CAS (physical abuse, emotional abuse, severe combined abuse and harassment) were identified from the analysis. They were not immediately obvious from the data but were inferred from how individual items grouped together. The Composite Abuse Scale was endorsed by the National Centre for Injury Prevention and Control,[105] as it has demonstrated reliability and validity for measuring IPV.

1.3.5. Integrating research paradigms

The diagnostic accuracy and validation paradigms represent different research methodologies for evaluating tests, including questions. As with the bringing together of the categorical and dimensional models in multidimensional scaling much can be gained by integrating the diagnostic accuracy and validation paradigms.

Though I have given examples of health research studies above that have used the diagnostic accuracy paradigm (see section 1.3.2.) and others that have used the validation paradigm (see section 1.3.3.), fewer studies have integrated and used methods from both paradigms. A study evaluating the PHQ did use both diagnostic accuracy indices and made limited use of the validation paradigm by employing kappa to show the agreement between diagnoses of depression made by the PHQ and those made by independent health professionals.[85]

IPV identification is unusual in that it is a topic specific research area in which researchers have used both the diagnostic accuracy paradigm and the validation paradigm whilst trying to find questions which identify IPV accurately. The evaluation of questions used to identify IPV in specific ethnic groups particularly benefits from these two perspectives. These methodologies are neither conflicting or contradictory but instead both help to bring us closer to developing questions for use in clinical histories that are both psychometrically robust and clinically useful; either able to identify between those with and without IPV or more pragmatically revealing how good questions are at identifying between the two. Many of these questions identifying IPV are already being used by health care professionals when they take clinical histories. Knowing what these questions may be measuring in women from different ethnic groups will aid the work of and decisions made by health care professionals.

1.3.6. A categorisation of validity evidence

The 1999 Standards for educational and psychological testing contain five categories of validity evidence,[106] as listed below:

A. Validity evidence based on the consequences of testing

B. Validity evidence based on relations to other variables

C. Validity evidence based on internal structure

D. Validity evidence based on response processes

E. Validity evidence based on test content

This comprehensive framework in effect encompasses and describes the many methods used in the validation paradigm. I have adapted these Standards by incorporating the diagnostic accuracy paradigm within them (in category B) as described in section 1.3.6.2.1.1. See figure 3, on page 50.

This integrative process informed my systematic review of questions trying to identify IPV in specific ethnic groups which covers research from both paradigms. It should be noted that the Standards focus on the process of construction of valid questions whereas my use of validity evidence is to aid my systematic review which appraises existing questions.

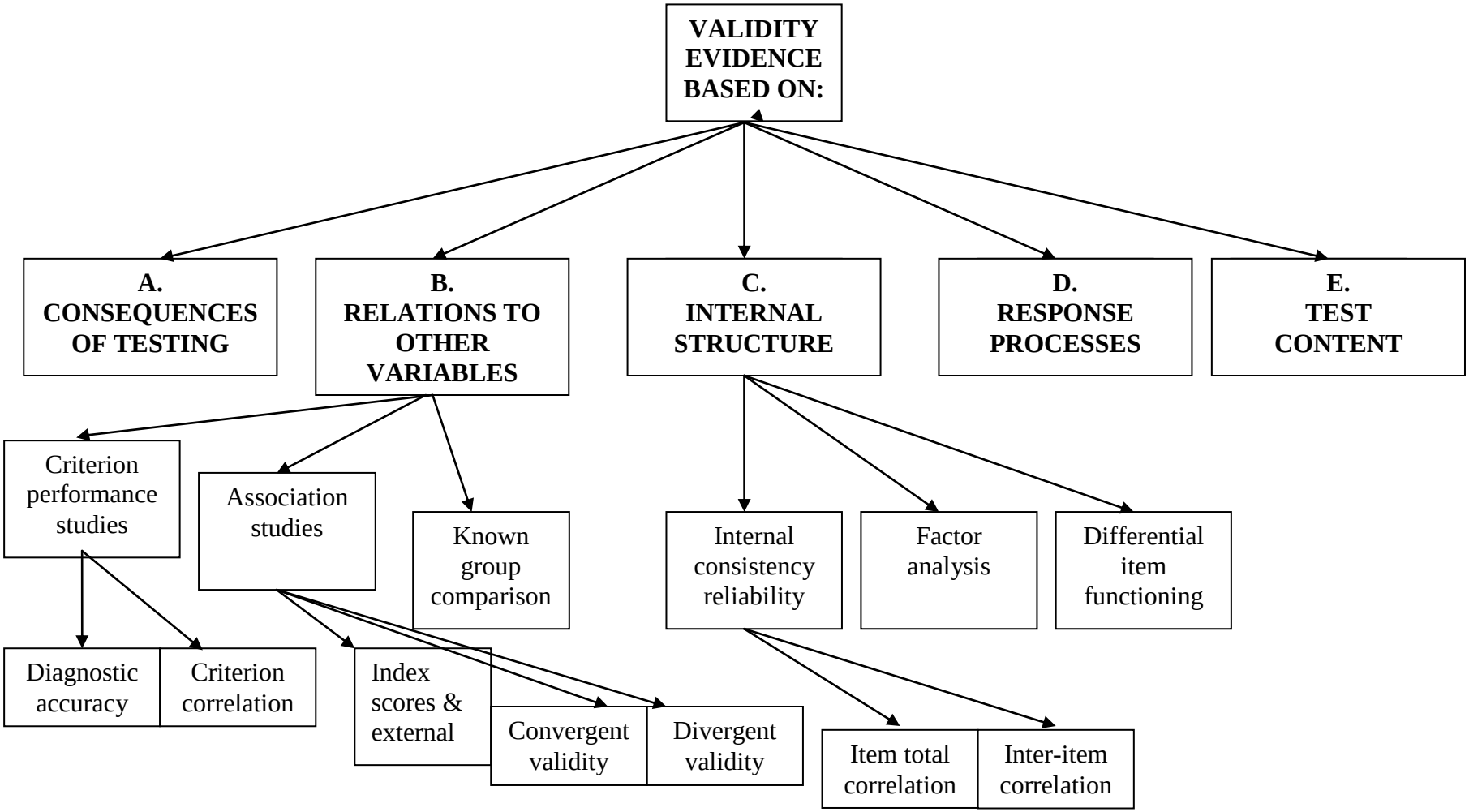


Figure 3: Categories and subcategories of validity

Throughout my account of the five integrated categories of validity evidence (see sections 1.3.6.1. to 1.3.6.3., 1.3.6.5 and 1.3.6.6.), I use the example of questions to identify IPV in different ethnic groups, to aid the understanding of specific categories. IPV identification is a research area that is facilitated by using this integrated framework.

The Standards for educational and psychological testing at time of publication (1999) were innovative as they put aside the longstanding traditional division of validity into what was and is known as the three Cs – of content, criterion and construct validity. In this "trinitarian" perspective, these three components of validity were considered to be relatively separate attributes which had to be independently established.[103] It was Anastasi who first contended that content, criterion and construct related validity did not relate to consequential individual test characteristics but were simply the derivatives of the developmental history of validation testing. He asserted that all validity was construct validity[107, 108] which included content and criterion-related validity. This was based on the principle that all test scores are based on constructs (see section 1.3.4.). Most recently, Streiner and Norman have concurred that validity is a unitary construct. They underscored this by drawing a clear distinction between validity and validation. They stated that validity refers to an outcome, (hence there are not different types of validity, for example content validity does not exist) whereas validation refers to the process of assessing validity for which there are many different types of testing (for example content validation does exist).[103]

Underlying what at first appears to be rather a semantic transformation in validity terminology, is an important principle that validation is a process which at its heart involves generating hypotheses which are then tested in a study. The study results should allow one to verify the degree of confidence one can place on inferences made about individuals on the basis of their score for a set of questions. Indeed the Standards highlight the interpretation of test scores by updating validity's definition to: "The degree to which evidence and theory support the interpretation of test scores entailed by proposed use of tests." Validation should be seen as an ongoing process which alters the degree of confidence that one draws about the inferences made.[81] The validity of a set of questions to identify IPV applies to the application of that set

of questions to a specific population not to the questions themselves. Hence it is not the questions (or investigations) that are valid but the uses of the questions that maybe valid.[89]

I will now describe the five integrated categories of validity evidence. I have also incorporated an additional sub-section (1.3.6.4.) on correlation which draws together the commonality of what initially appear to be quite different methods of measuring validity (criterion correlation studies, association studies and internal consistency reliability).

1.3.6.1. Category A: Validity evidence based on the consequences of testing

This "consequential validity" was introduced relatively recently to the Standards with continuing debate about its place in validation theory and practice. Accordingly there have been relatively few ideas about how to estimate consequential validity apart from descriptive studies addressing the extent to which anticipated benefits of measurement are realized using observations, interviews or other measures.[106] Crocker wondered whether validity evidence based on the consequences of testing should even be defined as an integral part of the validation plan, suggesting that it may then be seen as a socio-political process as opposed to being scientific and empirical.[109]

In contrast in the field of diagnostic accuracy, it is well established that index questions may discriminate well between those who have and do not have the target condition (category B evidence) but still do not necessarily affect the management of a condition.[110] Trials evaluating the clinical impact of the diagnostic strategy are then ideally needed.[94] Their evidence help health care professionals make good decisions about patient management based on tests that inform management that improves patient outcome as well as identifying a target condition. I consider that diagnostic strategy impact studies generate validity evidence based on the consequences of testing (category A).

In medicine, it is understood that medical tests ideally need to be evaluated in high quality studies prior to their dissemination and implementation in regular clinical practice.[3] A variety of study designs are used for this task including the diagnostic randomised controlled trial, before-after studies, cohort studies and case-control studies with the first deemed to be the most robust. Before-after studies can potentially be much quicker, are rooted in normal care and are an alternative if a randomised controlled trial is unfeasible or unethical.

1.3.6.2. Category B: Validity evidence based on relations to other variables

There are a variety of study types which generate validity evidence based on relations linking a test score to other variables. These can be organised into the three subcategories described below. The main difference between these subcategories and those listed in the Standards for Educational and Psychological Testing is that I have added the classification of "Association studies," as used by Rutjes et al. [3] (see figure 3, on page 29).

1.3.6.2.1. Criterion performance studies

Criterion performance studies (the first subcategory) includes diagnostic accuracy studies and criterion correlation studies. These studies explore the extent to which scores forecast or predict criterion performance. Criterion-related validity indicates the effectiveness of a test in predicting an individual's specified performance or report. These studies involve some type of comparison between the test score and the criterion. The criterion is a single empirical measure of the construct under study (for example IPV). It is equivalent to the reference standard, i.e. the best existing method at determining whether the target condition is present or absent. The standard experimental design for criterion performance studies is correlation (see criterion correlation studies, section 1.3.6.2.1.2.). This is probably as most measures are treated as being dimensional and not categorical. This is partly related to most measurement

tools using a dimensional scale with continuous judgements, as opposed to a categorical scale with categorical judgements. Psychometricians also tend to treat constructs as being continua.[103]

However criterion performance studies could be a classical diagnostic accuracy study design, if the construct is categorical and not only dimensional. Diagnostic accuracy is also a type of criterion-related validity where the reference standard provides the criterion against which the index test is validated.[3] Streiner also draws attention to the fact that though traditionally criterion validity has been assessed using a correlational study, it could also be assessed by a diagnostic accuracy study whereby a 2 X 2 table is used to calculate sensitivity and specificity indices as opposed to generating a measure of correlation from the 2 X 2 table such as the kappa coefficient (see section 1.3.6.4.2.).[81]

There are two types of criterion validity - concurrent and predictive validity. Concurrent validity is when a new scale and criterion measure are given at the same time and correlated. The criterion measure must be available at the time of testing. This methodology is most often used either when a shorter, simpler, cheaper or less invasive test is trying to replace a longer, more complex, expensive or invasive test. In research into identifying IPV it is when the scores on a new shorter set of questions trying to identify IPV are correlated with a criterion measure of IPV (a longer set of questions). Questions that not only predict the criterion but that can additionally show the changes in sub-scales responsible for the criterion changing maybe more useful than the present criterion measure.

Predictive validity is when a new scale generates answers, including identification, earlier than the current criterion measure. Hence the criterion measure result is not available until some time in the future (this may be days or years later), after the new scale has been administered. For example, a diagnostic test may have to await disease progression to either confirm or reject its predictions.

1.3.6.2.1.1. Diagnostic accuracy studies

As mentioned earlier, the classical diagnostic accuracy paradigm compares the results of index questions to a standard reference. The index questions should ideally be able to identify whether a target condition (for example IPV) is present or absent but their ability to do so is not known. Diagnostic accuracy studies all try to measure the accuracy of the index questions at identifying the target condition, by assessing the degree of agreement between the results of the index questions and the results of the reference standard. This established approach is represented in figures 4(a) and 4(b)[3] on page 56. Accuracy is a phrase originating from measurement theory. It is the closeness of agreement between an analytical measurement and its actual true value.[111]

In diagnostic accuracy studies one would normally recruit a group of individuals some of whom are potentially affected by the target condition whilst some are not. Firstly the index questions would be administered to all the participants and would generate an index score showing whether according to the index questions that the target condition is present or absent. Following this the reference standard would be administered to all participants. This would indicate whether according to it that the target condition is present or absent. Figure 4(a) illustrates this classical design of a diagnostic accuracy study. This then allows the results of the diagnostic accuracy study to be compiled in a 2 by 2 table, as shown in figure 4(b). The perfect diagnostic accuracy study would have a faultless reference standard identifying the target condition without errors, all index scores would be compared to the same reference standard with both being administered at the same time.

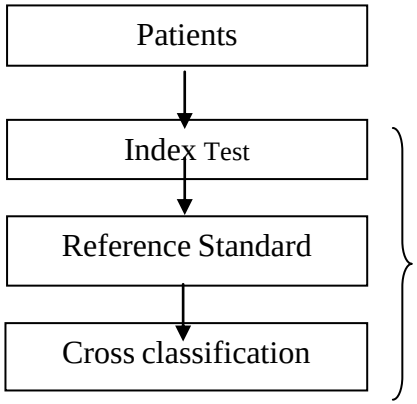


Figure 4(a): Classical design of a diagnostic accuracy study

Target Condition

Index Test + / Result -	**Present**	**Absent**
+	TP	FP
-	FN	TN

Accuracy measures:

Sensitivity = TP/(TP + FN)
Specificity = TN/(TN + FP)
PPV = TP/(TP + FP)
NPV = TN/(TN + FN)
LR+ = [TP/(TP + FN)]/[FP/(FP + TN)]
LR- = [FN/(FN + TP)]/[TN/(TN + FP)]

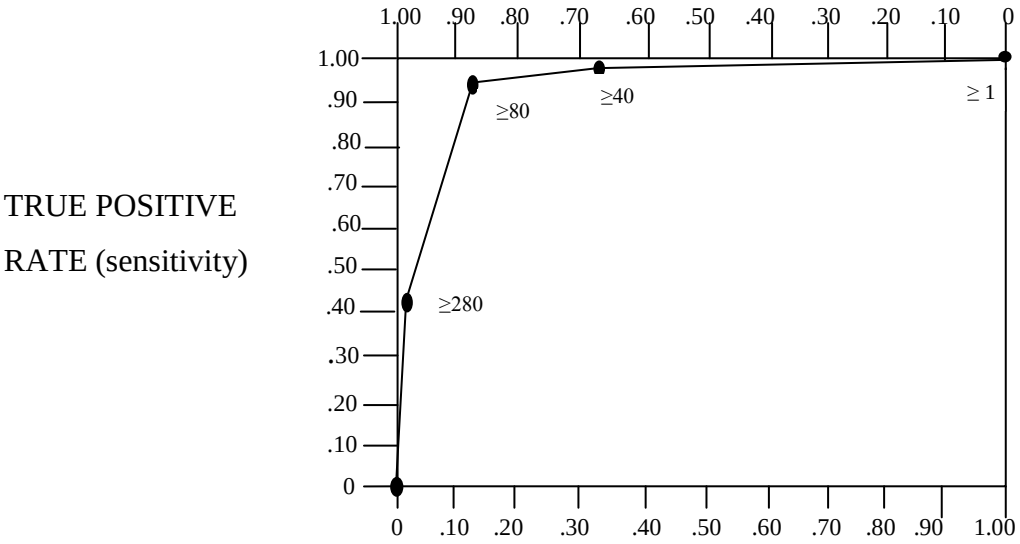
Figure 4(b): Results of an accuracy study in the case of a dichotomous index test result

TP, true positive result; FP, false positive result; FN, false negative result; TN, true negative result; PPV, positive predictive value; NPV, negative predictive value; LR, likelihood ratio.

I will now define and consider the various diagnostic accuracy indices. A relatively detailed explanation has been provided as this terminology frequently appears in IPV identification papers and this thesis. Despite the regular use of these terms in the literature, there are common misconceptions about some of these terms, including that the most important characteristics of a test are its sensitivity or specificity. Figure 4(b) lists the mathematical formulae used to calculate these diagnostic indices.

Sensitivity refers to how good the index questions are at picking up people who have the target condition.
Specificity refers to how good the index questions are at correctly excluding people without the target condition.
Positive predictive value (PPV) informs us if a person tests positive, the probability that she has the target condition. It is also known as the post-test probability of a positive test.
Negative predictive value (NPV) informs us if a person tests negative, the probability that she does not have the target condition. It is also known as the post-test probability of a negative test.[112]
The likelihood ratio (LR) of a positive test is how much more likely is a positive result to be found in a person with, as opposed to without, the condition.[112] Their advantage over predictive values is that they are more constant with prevalence changes.
The post-test odds (PTO) permit the background prevalence to be factored into the LR.
The receiver operator characteristic (ROC) curve is constructed by plotting the sensitivity of each individual score against its false positive rate (= 100 – specificity) See figure 5, on page 58 which uses the example of creatinine kinase values in myocardial infarction. The ROC curve can be used to determine the optimal cut off score which maximises the true positives whilst minimising the false positives (i.e. the point that has the highest combined sensitivity and specificity, in the top left hand corner of the ROC curve). The **area under the ROC curve** measures the performance of a test.[113] Its value can lie between 0.5 (i.e. test has a likelihood ratio of 1 for all its cut-off values and so is unhelpful) and 1 (i.e. test perfectly separates those who have the target condition from those who don't).

SPECIFITY



TRUE POSITIVE RATE (sensitivity)

FALSE POSITIVE RATE

Figure 5: An ROC curve for creatinine kinase values in myocardial infarction[61]

For some tests and target conditions maximising the true positives (i.e. maximising sensitivity) and minimising the false positives (i.e. maximising specificity) may not be the most important guiding principal determining the cut-off score. If a false positive result was very damaging (for example committed a patient to an invasive test) one may choose a different cut off point which minimises the false positive rate (i.e. maximises the specificity by being towards the left hand side of the curve). Whereas if a false negative result was very hazardous (for example missing underlying aggressive cancer) one may pick a cut off point that maximised the true positive rate (i.e. maximises the sensitivity by being higher on the curve).

This analysis of the use of sensitivity and specificity in determining cut off points should not detract from that which is more clinically and practically informative about index questions, i.e. their predictive values. Sackett considered the relative functional importance of diagnostic indices and supported that the predictive values (PPV and

NPV) are far more instructive than the sensitivity and specificity of index questions in clinical practice.[61] The capability of the answers to index questions to change one's mind from what one thought before administering the questions (i.e. the pre-test probability of IPV, also known as the IPV prevalence) to what one thinks afterwards (i.e. the post-test probability of IPV, also known as the PPV) is important. If there is a large change from pre-test to post-test probability, the index questions are likely to be very useful in real clinical practice.[91] Whereas the sensitivity and specificity interpret the index questions' results retrospectively, it is the PPV and NPVs that actually establish the predictive properties of the index questions in the future. Hence Sackett argues that when tests are used clinically one does not know who has and does not have the target condition. Thus the predictive values of a test (i.e. PPV and NPV) are much more useful clinically.[61] However these predictive values can only be interpreted alongside prevalence, as invariably predictive values always vary with prevalence. As prevalence decreases, PPV decreases with it and NPV increases. Therefore even brilliant index questions that have a sensitivity and specificity of more than 95% may have a rapidly decreasing PPV as the prevalence falls and so clinically be poor index questions which produce no substantial difference between the pre-test to post-test probability.[61]

With the pre-test probability (prevalence) of IPV having a very wide variation around the world (see section 1.2.1.1.) it is difficult to be prescriptive about laying down pre-established criteria to classify low, moderate or high predictive values for index questions trying to identify IPV. The most clinically useful index questions may be those that are found to produce the largest difference between the specific study prevalence and the PPV of the index questions in that study.

1.3.6.2.1.2. Criterion correlation studies

These studies involve correlation between a predictor index test score (of the new set of questions) and the criterion score. This forms the criterion-related validity coefficient which assesses the validity of the index questions. The criterion-related validity coefficient has a number of characteristics in common with other correlation

coefficients which I will be covering in section 1.3.6.4., including the different statistical methods used to calculate them.

1.3.6.2.2. Association studies

Criterion performance studies require a criterion measure or a reference standard. If no reference standard exists (which could be argued is the case with IPV) then an association study may need to be undertaken. Association studies include those that look at the association between index scores and external variables, convergent validity studies (looking at the association between two tests measuring similar constructs) and divergent validity studies (looking at the association between two tests measuring dissimilar constructs). In association studies it is the underlying theories about the target condition (IPV) which generates hypotheses regarding potential associations, expressed quantitatively, between index questions and attributes that can be evaluated. For example, one may hypothesise that a woman experiencing IPV will visit a doctor more often or have worse mental health. If the theory is erroneous, the quantitative association may be deceptive. If the hypothesised association is not seen between questions trying to identify IPV and other observations (be it an external variable or another test score), one has to decide whether the index questions have low validity or that the theory is incorrect or both.

1.3.6.2.2.1. Between index scores and external variables

Association studies measuring correlations of the type and extent of relationships between index scores and external variables for example hospitalisation, use of services, readmission etc. can evaluate the capacity of a set of questions trying to identify IPV to correlate with external variables.

1.3.6.2.2.2. Convergent validity studies

These studies investigate the relationships between index scores and other tests intended to measure similar constructs. Neither test is purported to be a reference

standard. However the two should be related, according to the theoretical basis of the construct and so should correlate moderately highly. If the correlations between the two tests are too high this implies that the new test is unnecessary replication unless it has some advantage over the older test (for example is shorter). If the scores on the two sets of questions do not correlate, this would point to either a difficulty with the new set of questions (for example they are not identifying IPV) or an issue with the theory linking the two sets of questions (for example do the two sets of questions identify different dimensions of IPV). There would be no way of knowing from the results of the correlation alone.

In comparison with criterion performance studies (this includes diagnostic accuracy studies and criterion correlation studies), convergent validity is based on more assumptions and so is a less robust method. Consequently in my framework, convergent validity appears below criterion-related validity which has fewer underlying assumptions (see figure 3, on page 50).

1.3.6.2.2.3. Divergent validity studies

Divergent validity studies: (also known as discriminant validity studies) are closely related to convergent validity studies but investigate the relationships between scores and other measures of different constructs. These relationships should not correlate or have low correlations.

1.3.6.2.3. Known group comparisons

Known-group comparison studies are intended to test hypotheses about expected differences in test scores across specific groups of examinees. Study populations hypothesized to differ on a test construct (for example non-abused women compared with abused women living in a refuge) are assessed using the set of questions trying to identify IPV. If the expected mean differences in scores are found, the questions identifying IPV are supported. Streiner and Norman also refer to this as construct validation by extreme groups or discriminative (not discriminant) validity.[81] Known

group comparisons contain an inherent study population bias as ambiguous cases will have been eliminated. However they sometimes are a necessary study design when a reference standard does not exist.

1.3.6.3. Category C: Validity evidence based on internal structure

The role of validity evidence based on internal structure is constrained in that it does not actually tell us what the questions are identifying and whether they are identifying what is required to be identified. To achieve this, data external to the questions is required (i.e. category A and B).

1.3.6.3.1. What is internal consistency reliability?

For a set of questions to effectively identify a construct requires the individual questions in that set, to also represent the same construct, assuming that the construct is one-dimensional. Answers on questions should be moderately correlated with each other (inter-item correlation) and each individual question's score should correlate with the total scale score (item-total correlation). This would reflect a high degree of homogeneity also known as the internal consistency of the scale which is a type of reliability measure. This represents validity evidence based on the internal structure of the set of questions.[89] Internal consistency can be measured easily by simply administering the questions once to participants as it is generated by looking at the average of the correlations amongst all the questions in the group. This advantage over other reliability coefficients (see section 1.3.6.3.4.) means that internal consistency reliability coefficients are commonly seen in published papers.

1.3.6.3.2. Internal consistency and classical test theory

Streiner and Norman advise that in most situations, when measuring a construct the set of questions measuring it should be homogeneous.[81] The theoretical assumption that if questions are highly correlated, the construct of interest has been measured to some degree of consistency[114] arises from Classical Test Theory. In Classical Test Theory, any test is constructed from a random sample of all the possible questions that could be in the test. There is a supposition that there is a "universe" of questions that identify a given trait or behaviour; and that a scale is made up of a random subset of these questions. Therefore the questions should be highly interrelated if they are assessing the same construct and the scores would be reliable.

1.3.6.3.3. When should internal consistency not be considered

Homogeneity should not be measured across different subscales when questionnaires are multidimensional.[81] Factor analysis, another type of validity evidence based on internal structure, evaluates whether individual questions belong to different dimensions.

Factor analysis has been used in the development of long research tools not used in routine clinical practice to identify IPV. For example, the 30 item Composite Abuse Scale has been shown to measure four dimensions of abuse inflicted on a woman by her partner. These dimensions are physical abuse, emotional abuse, severe combined abuse (which includes sexual IPV) and harassment.[101, 115] When the Composite Abuse Scale is presented to women the questions from each dimension are mixed together randomly. It would be illogical to measure the internal consistency of the 30 questions that refer to different dimensions of IPV together. Instead it is the internal consistency of the individual dimensions that has been calculated.

Therefore for new sets of questions trying to identify IPV one needs to decide whether the questions operate over one dimension or are multidimensional before deciding whether internal consistency estimates have any role in providing validity evidence.

For a set of questions that have good content validity (see section 1.3.6.6.) capturing physical IPV, sexual IPV and emotional IPV, the questions address different dimensions of IPV so internal consistency measures should not be applied across them. Some sets of questions trying to identify IPV may only focus on one dimension of IPV, for example physical IPV. This would result in decreased content validity with regards to representing the whole spectrum of IPV but being uni-dimensional it would be methodologically correct to measure the internal consistency of these questions. Streiner advises that the aim of a scale is inferential which is more dependent on its content than its internal consistency, making the former more important.[103]

Additionally, if a test is trying to categorically divide women into different groups, then it may be that it contains questions which do not actually relate directly to the specific construct and the internal consistency of the set of questions becomes irrelevant.

1.3.6.3.4. Internal consistency in relation to other reliability measures

Reliability of a set of questions identifying IPV would be evidence that these questions were measuring IPV in a reproducible manner. Internal consistency is only one type of reliability measurement. The internal consistency reliability coefficient does not take into account other types of reliability representing other sources of variation (i.e. errors of measurement) such as that caused by different times of test administration (test – retest reliability coefficient), observer to observer variation (inter observer reliability coefficient) and variation by the same observer (intra observer reliability coefficient). The internal consistency reliability coefficient is completely independent from these other reliability coefficients.[81] Therefore it does not give the full picture of the true reliability for a group of questions identifying IPV. Internal consistency reliability for any measure can often be much more positive than the actual total reliability.

1.3.6.3.5. Statistical methods to calculate internal consistency reliability

Item-total correlation and inter-item correlation are the two principal methods used in a number of statistical tests to calculate the internal consistency for a set of questions.

1.3.6.3.5.1. Item-total correlation method

This is a frequently used method for examining the homogeneity of a group of questions. It involves checking the correlation of an individual question with the group of questions having omitted that question. Each individual question's score should correlate with the total scale score. A formula devised by Nunnally separates an individual question's contribution from the whole score.[116] Kline advised that an individual question should correlate with the total score above 0.2 whilst questions with lower correlations should be rejected.[117] The Pearson product –moment correlation coefficient is used if there are more than two response alternatives, even if data is not normally distributed (see section 1.3.6.4.1.). The point biserial correlation coefficient is used if questions have only two response alternatives (see section 1.3.6.4.3.).

1.3.6.3.5.2. Inter-item correlation method

A number of different statistical tests used to calculate internal consistency reliability use the inter-item correlation method, including the Kuder-Richardson formula 20 reliability coefficient (used when individual questions are scored dichotomously) and Cronbach's alpha (an extension of Kuder-Richardson 20 which can be utilized when there are more than two response choices).

Cronbach's alpha, also known as the Coefficient alpha or alpha, is the most widely used statistical method for calculating inter-item internal consistency probably as it can be used for both dichotomously scored questions and those with multiple response categories, for example the Likert scale. Unfortunately, there is a great deal of misunderstanding about Cronbach's alpha and what it actually means though it ubiquitously appears in papers developing scales. Central to understanding

Cronbach's alpha is that it not only represents the magnitude of correlations among questions but also the number of questions in a scale. Hence by simply doubling the number of questions, one increases Cronbach's alpha despite the average correlation remaining unchanged.[89]

There is also no absolute clear consensus regarding the ideal numerical value for Cronbach's alpha. Aaronson et al's commonly accepted minimal standards for reliability coefficients was 0.7 for group comparsions; and 0.9-0.95 for individual comparisons.[118] Nunnally agreed that alpha should be > 0.7 but that alpha should be no higher than 0.9 as this may imply that some questions were redundant, adding little extra information, as they make the same enquiry in slightly altered ways.[116] Streiner & Norman sum up that the real problem of having a Cronbach's alpha with a number between zero and one is that it does not lend itself to commonsense interpretations. They emphasise that high alpha values should always be interpreted with great caution and never assumed to be inherently good. They state that internal consistency should be greater than 0.8 with higher values depending on the use of the test and the cost of misinterpretation.[81]

1.3.6.4. Correlation and correlation coefficients

Correlation between two measures is used in a number of methods described above to measure validity, including criterion correlation studies, association studies (between index scores and external variables, convergent and divergent validity studies) as well as all reliability measures including internal consistency reliability. These correlation coefficients have some common features. All can be affected by the difference within the group being studied and test length. Hence the more heterogeneous and varied the study population, with a wider range of test scores, the larger the correlation coefficient; whilst the more homogeneous the group, with a narrow range of test scores, the smaller the correlation coefficient. Consequently a small correlation coefficient could be the result of strict sample selection causing a restriction in variance. This can be explored by actually inspecting the scatter plot showing the bivariate distribution between the test score and the other measure. This can also

clearly illustrate whether the relationship is linear and uniform; or if the two variables are related non-linearly but have zero correlation. Different correlation coefficients are defined by various statistical methods which I have described below.

1.3.6.4.1. Pearson's product moment correlation coefficient

This is also known as the Pearson correlation, the Pearson's correlation coefficient, the Pearson product moment correlation and the r value, denoted by "r." This reveals how close the relationship between two measures can be described by a straight regression line.[112] Both measures need to be continuous but do not need to have the same units. The correlation coefficient is the sum of products divided by the square roots of the sums of squares of X and Y and hence has no units. This also makes the correlation coefficient lie between -1.0 and +1.0 which relates the closeness of the linear relationship between the two measures. If the relationship is strong the correlation coefficient approaches +1, if it is weak it moves towards 0 whilst if there is a negative relationship (i.e. if one measure goes up, the other goes down) it would be closer to -1.

The product moment correlation coefficient presumes equal variability throughout the range of scores. This is exhibited on a scatter-plot. It is the best coefficient to use in almost all cases if there are more than 2 response alternatives. The product-moment correlation is robust enough to produce relatively accurate results, even if data are not normally distributed.

1.3.6.4.2. Kappa coefficient

This correlation coefficient is also known as Cohen's kappa statistic [119], kappa statistic and just kappa. It measures the correlation between two dichotomised measures (for example the presence or absence of IPV, alluding to the categorical model) as opposed to two continuous measures (for example the degree of IPV present, alluding to the dimensional model) when using the product moment

correlation coefficient. This approach calculates simple agreement, i.e. the proportion of responses in which the two observations agreed. The kappa coefficient represents the proportion of responses in the two agreement cells (yes / yes, no / no) in relation to the proportion of responses in these cells which would be expected by chance, given the marginal distributions. Therefore it demonstrates the degree of agreement which has occurred over and above that which would have occurred by chance alone. Its weakness is that it is influenced by the average prevalence of the target condition (i.e. IPV).

1.3.6.4.3. Point biserial correlation coefficient

The point biserial correlation coefficient[120] is mathematically equivalent to the Pearson product moment correlation but is used if there is one continuously measured variable and a dichotomous variable.

1.3.6.4.4. Spearman's correlation coefficient

This assesses the strength of the association between two continuous variables when one cannot assume that the data were sampled from a particular type of distribution (i.e. a non-parametric statistical test).[112]

1.3.6.5. Category D: Validity evidence based on response processes

Evidence of validity based on response processes considers the ways in which individuals respond when completing test questions. Therefore when trying to decide whether questions measuring a construct are valid, one needs to also consider whether answering the question generates tasks that require complex activities which actually

impede the measuring of that construct. Some questions may cause construct-irrelevant variance due to the way that they are asked.[121]

When constructing new questions to identify IPV they must be easily interpretable. Words that are ambiguous, incomprehensible or contain jargon terms only used by professional groups need to be eliminated. The reading level of the questions generally should be no higher than the reading age of a 12 year old.[103] The response alternatives should be precise especially with regards to time. Badly selected vague questions with poor wording cannot be overcome by complex statistical analysis. Validity evidence based on response processes becomes especially important when using questions in different cultural groups which may be less familiar with for example, a Likert type scale [122] or a "true / false" format.

Likert scales have a number of response alternatives with interval properties. The participant has to rate them according to her degree of agreement or disagreement. Likert scales are thought to be difficult to negotiate for diverse patients with poor literacy skills,[122] decreasing the value of the validity evidence generated based on response processes.

1.3.6.6. Category E: Validity evidence based on test content

The content validation of the index questions measuring a construct is a systematic analysis of the appropriateness of the questions' content. The face validation is a subjective judgement to see whether on the face of it the questions appear to be assessing the same construct.[81]

Content validation refers to more than just the mere content of questions in that it also concerns itself with the range of responses generated by the content of the questions. Hence the range of responses that are elicited should represent the complete domain that one is trying to measure. This involves carefully specifying the entirety of behaviours that can occur in that domain.[82]

Therefore questions that try to identify IPV should arise from a comprehensive definition of IPV so that they capture the different types of IPV. The WHO has globally defined intimate partner violence (IPV) as:

"Any behaviour within an intimate relationship that causes physical, psychological, or sexual harm to those in a relationship; it includes: physical aggression, psychological abuse, forced intercourse and other forms of sexual coercion, various controlling behaviours."[123]

Hence a set of questions trying to identify IPV needs to include questions on physical, sexual and emotional IPV otherwise the questions cannot identify it. This flaw cannot be corrected by statistical manipulation. Each separate question should focus on one element of IPV only. One question cannot cover more than one type of IPV as this would result in the question being too complex and difficult to answer using the common "Yes / No" answer. Different patterns of IPV can occur so no assumptions should be made about whether physical, sexual or emotional IPV are either mutually inclusive or exclusive.[17]

Fontes states that researchers may use definitions that do not exist in other cultures.[124] Lachs affirms that as there is no universally agreed case definition of IPV, one cannot calculate the sensitivity or specificity of IPV tools.[125] Undoubtedly IPV does mean different things to different women due to its opaque nature. This may be more accentuated in women from different ethnic groups (see section 1.5.1.) though may be just as evident in women from the same ethnic group. In clinical consultations it is the individual woman who decides whether to characterize her experiences as IPV. This interpretative process may be accommodated by the health care professional giving out key messages over time such as the actions by her partner, for example marital rape, are illegal; that there is no excuse for this behaviour and that this behaviour cannot be the woman's fault. I do not think that the woman's role in describing her own experience of IPV contradicts the importance of a global definition of IPV.

It may be difficult to define IPV globally but the WHO definition (see above) is the closest to a universally agreed definition. Hence if a set of questions is going to be

used to identify IPV in a variety of settings it is important to use the WHO definition to assess content validation as opposed to colloquial definitions.

Though IPV may be interpreted differently around the world, some women even construing it to be "normal," international human rights law is very clear that states have a duty to prevent, prosecute and punish violence against women.[6] Using a human rights framework in which aspiration to health equality[126] and violence against women are human rights issues[127] leads to an appreciation that cultural relativism should not be used to diminish violence and its effects.[128], [129] Whereas the culture of a group is a constantly changing phenomenon[130] (see section 1.4.1.3.), the IPV definition arising from a human rights framework should be a constant.

1.3.6.6.1. Translation of questions

Translation may significantly alter the meaning of questions unless attention is paid to reassessing the content validation of translated questions. The translation of questions to identify IPV also highlights the opacity of IPV as a construct. The content validitation of translated questions requires consideration of conceptual equivalence, item equivalence, semantic equivalence, operational equivalence, measurement equivalence and back-translation.[81]

The first and most important step, yet also the most difficult to achieve is conceptual equivalence. This means establishing whether the persons in the two language groups which represent two cultural groups actually perceive the concept in the same way. This may be most difficult with aspects of emotional IPV.

Item equivalence involves checking that specific questions are relevant and acceptable in the new language group. Semantic equivalence then checks if the meaning of each word is equivalent in the two language groups. For example, the direct translation of "I feel blue" from English to Spanish may not be semantically equivalent as there maybe no association of blueness with sadness in Spain.

After conceptual, item and semantic equivalence have been completed, one can proceed to the actual translation task. Ideally, there should be at least two separate translations preferably within two teams, translating into their native languages, using colloquial language that participants are more likely to be familiar with rather than the more formalised speech often used by and between professionals.[131] Independent teams who are preferably unaware of the research objectives and what the instrument is measuring, and have not seen the original English questions that were being translated should back-translate each question into English. Finally a separate team of translators should ideally look at the original and back-translated versions to resolve any outstanding differences.[103]

Operational equivalence is considering whether the same questions, with the same instructions and the same method of administration (be it a self-completed / telephone / mailed questionnaire, face-to-face interview, computer assisted administration) would function effectively in the new language group. For example, many first generation female Bangladeshi migrants to east London may not speak English, instead speaking Bengali but they would not necessarily be able to read Bengali. The literacy rate in Bangladesh is lower compared to the UK. Therefore they may find it difficult to answer a self completed translated questionnaire even though it is in Bengali.

Once one has a translated version of the instrument one cannot assume that it has the same reliability and validity as the original version so these characteristics need to be reassessed in the new tool. One could then proceed to see if any cut-off scores used in the earlier tool are suitable in the new translated version. Once one has two versions of questions in two languages, the differences in the results (and even the similarities) need careful interpretation. Differences may not just be due to ethnic or cultural differences between two groups but also due to other factors. For example, socio-economic differences would need to be carefully examined. Interpreting similarities or differences requires care and complex analysis.[103]

Having examined the background to the measurement of validity, I now consider the background to the term ethnicity and how it has been used in IPV research.

1.4. Ethnicity

Throughout this section on ethnicity, I will illustrate key issues by using IPV research studies which have used ethnicity data whilst considering how ethnicity may impact on IPV identification. In this section I will firstly explore the concept of ethnicity and then how ethnicity is actually utilised in health research studies. I use the phrase "health research" as an umbrella term which includes epidemiological, clinical, and health services research. Looking at how ethnicity is actually used in studies involves exploring ethnicity's close relationship with race. I will then explain why I have chosen to study ethnicity. I will focus on the rationale for and potential dangers of collecting ethnicity data in research studies before lastly presenting five criteria to assess the use of ethnicity data by papers.

1.4.1. What is an ethnic group?

Ethnicity is derived from the Greek work "ethnos" which means a nation, people or tribe. The Oxford Dictionary[90] contains a variety of definitions for the adjective ethnic including:

"relating to race or culture (ethnic group);
(of a social group) having a common national or cultural tradition."

Mares and colleagues,[132] stated that an ethnic group does not need to be stringently demarcated by specific cultural factors or characteristics but rather that the important feature of an ethnic group is that it is recognised by its own members and by others. Hastrup[133] said that

".. meaning of ethnicity cannot be sought out in a purely deductive manner,"

– by others, be it one's patients or ethnicity experts –

"it requires the cooperation of the people involved ... they themselves play the part of theoreticians in this field…"

This highlights the importance of self identity within the realm of ethnic classification. For ethnic classifications to be valid in studies they need to be self assigned as opposed to being determined by others.[134]

Anthropologists have viewed ethnicity as a result of interaction, rather than representing the innate characteristics of a human group.[135, 136] Ethnicity means different things in different contexts.[137] Cohen points out that ethnicity along with age, gender, class and other characteristics that individuals use to define them-selves all have an objective status derived from economic and social realities; and a subjective status with a symbolic quality. This therefore allows the simultaneous expression of both individual and collective identities.[138]

An ethnic group has been defined by shared characteristics including cultural traditions, languages, religion, ancestral and geographical heritage.[134, 139] A comprehensive definition of an ethnic group is

> "a collectivity within a larger society having real or putative common ancestry, memories of a shared historical past, and a cultural focus on one or more symbolic elements defined as the epitome of their peoplehood.[140]"

In the context of investigating questions designed to identify IPV, the foremost symbolic elements may be cultural beliefs about gender, family and what constitutes abuse. It is these cultural beliefs that impact on an individual's ethnic identity and ethnicity (see section 1.4.1.3.).

1.4.1.1. Ethnicity's relationship to race

Despite the wide ranging debate on what constitutes an ethnic group, health research generally uses a narrow concept of ethnicity in studies.[137] Ethnicity is rarely based on any cultural factors or cultural beliefs. Previously in the UK, the country of birth was often used as a proxy measure for ethnicity. In the 21st century, in increasingly diverse populations, country of birth has become a poor indicator of ethnicity. Instead now in Western Europe and the US, ethnicity is often based on apparent racial

categories, the majority of which represent the degree of melanisation of skin cells, i.e. skin colour. African-Americans equates to black, Asian to brown and white to white. I consider that the main justification for basing classification on euphemisms for skin colour (apart for the few conditions that are directly related to the degree of melanisation, for example melanoma, vitamin D deficiency) is that this could be used to investigate the effect of racism on health. Disappointingly researchers using these classification systems rarely mention racism whilst at the same time failing to reiterate that neither ethnic or racial groups relate to either biological or genetic differences.[141]

Geneticists have clearly shown that the genetic differences between so-called races is less than the differences seen within these groups.[142] The principle that ethnicity and race are social constructs rather than biologically based ones is supported by a number of professional organisations, including the American Academy of Pediatrics,[143] the US Surgeon General,[144] the American Psychological Association,[145] the American Sociological Association [146] and the American Anthropological Association.[147] It has been said that the race concept has gradually changed, incorporating shared histories and social factors, hence merging with the concept of ethnicity.[148]

However I consider that the current use of both ethnicity and race in health research, at the grassroots level reflected in academic papers does not represent a convergence of the two constructs. Even though there is little practical difference in the way that the two terms are currently used, this does not symbolize a union. Ethnicity has essentially evolved into becoming a politically correct way of saying race. Authors feel more comfortable and safe using the term ethnicity as opposed to race. Consequently, this has resulted in the burden of ethnicity classifications often being precisely the same as that for racial classifications. Ethnicity data is as vulnerable to discriminatory or prejudiced interpretation as is racial classification data to racist interpretation. Any research that contains ethnicity or race data is susceptible to stereotyping and discrimination. Baldwin agreed that ethnicity was similarly tainted to race, both ethnicity and race being derived partly from immigration and hence politically loaded terms as opposed to being neutral.[149]

1.4.1.2. The phrase “race / ethnicity”

In health research literature, as well as the trend whereby ethnicity often simply replaces the word race, there is also currently the widespread use of the phrase “Race / ethnicity” – reflecting the synonymous use of race and ethnicity in practice.[134, 137, 148, 150] The use of the phrase “Race / ethnicity” has been supported by some:

“Race / ethnicity: in which race can be considered “the category to which others assign individuals on the basis of physical characteristics, such as skin color or hair type, & the generalizations & stereotypes made as a result” and ethnicity as “group mores & practices of one’s culture of origin.”[145]

I think that the use of race / ethnicity reflects the similarity of the current ethnic and racial classifications that are used, as discussed above. I have seen no clear justification for combining these two quite different concepts into one term. Instead, I consider that the existence of the phrase “race / ethnicity” results in continued use of terminology which can purport to be ethnicity though is based on race which in turn largely reflects skin colour.

Despite the current similarity in how ethnicity and race are practically used in health research literature, they are fundamentally different. In the future, ethnicity may potentially throw light onto the complex issue of identity, unlike race. Ethnicity’s strength is that when using it one can also include the factors that describe ethnicity. Therefore an understanding of ethnicity in health research still has the potential to be developed unlike race. Bhopal also stated that ethnicity is still under development whilst having the capability to combine cultural, social and biological features.[134] Oppenheimer concurred that though ethnicity has its own load of political, social and ideological meaning, being closely aligned to culture it is preferable to race.[151]

1.4.1.3. Studying ethnicity

The decision on whether ethnicity or race is studied should be determined by the specific research question that one is trying to answer. Investigating ethnicity is not always preferable to race. For example, it may be appropriate to investigate race, if the research question hinges on inequity driven by racial appearance. If the research question centres on behaviours, cultural beliefs and identity then it may be ethnicity which is more appropriate.

In the context of investigating questions trying to identify IPV, I think that the cultural beliefs that impact on an individual's ethnic identity and ethnicity are central. Cultural beliefs are best uncovered by talking to individuals who are the experts on the cultural factors at play in their lives.[152] Cultural beliefs are likely to be particularly important when considering how individuals interpret IPV, in that cultural beliefs will impact on gender roles, expectations about family roles, what is considered to be abusive, how willing an individual is to disclose abuse and the reasons why they would consider disclosing abuse. The premise for my thesis is that cultural differences in attitudes towards IPV could affect disclosure in different ethnic groups which in turn could also affect how accurately some questions (according to their precise wording and order of words) identify IPV in different ethnic groups in a health setting. It is not only culture that may affect the questions used to identify IPV but also women's experience of IPV that may affect their culture. The culture of a group constantly changes as a result of the people in that group engaging and reinterpreting it. A postmodern perspective widely accepts the changeability of culture: "culture is not a static phenomenon; individuals interact with their culture so that the culture is constantly challenged and redefined."[130]

My thesis is concerned with the identification of IPV which may be impacted upon by cultural beliefs as well as cultural position (i.e. minority verses majority communities, oppressed groups verses oppressors - also see section 1.4.2.1.). Hence it seems appropriate that my thesis should focus on differences in ethnic groups as opposed to racial groups. Whilst reviewing previous literature (both in this background chapter and in my systematic review) I have used the ethnicity terms and factors used to

describe ethnicity reported in those studies. I now consider the rationale and dangers of collecting ethnicity data in general in health research.

1.4.2. Rationale for collecting ethnicity data in health research

The rationale for collecting ethnicity data in research studies is to expose health inequalities, improve health and to respond to increasing ethnic diversity.

1.4.2.1. Exposing health inequalities

The study of ethnicity can help to expose inequalities in both health and healthcare. Ethnicity as well as race, socioeconomic status, education level, health behaviours, gender, age and occupation are all well recognised epidemiological exposure variables.[134, 153] They can be used to subdivide populations, showing differences in disease experience. They all define a possible group identity, helping to make inequality more meaningful in a given population[154] as they are markers of underlying factors which are relatively more difficult to measure, such as how powerful an individual is and the power dynamics of relationships that they have with others around them (for example, intimate partners, extended family, neighbours, employers or the state).

Epstein visualises ethnicity as well as race, age, gender, social class and sexual identity as "…intersecting attributes of identity, markers of difference, dimensions of social hierarchy and power…".[140] Consideration of power relationships is particularly important when reflecting on IPV and how IPV maybe interpreted differently in specific ethnic groups. Power differences between the sexes may increase the likelihood of gendered violence whilst power differences between ethnic groups raise the possibility of cultural violence (including racism and discrimination). The power relationships between ethnic groups are often glossed over.[155] How violence operates to link the power relations of gender and ethnicity has been

considered by sociologists.[156] Collins describes that both hierarchies of gender and ethnicity as well as race, class, age, nation and sexuality are supported by violence. For example, gender hierarchies are supported by pornographic images of women, workplace sexual inequalities, widespread physical and sexual IPV. Collins advises that using a gender-only framework (or an ethnicity-only framework) restricts the understanding of an African-American woman's experiences with violence. In effect IPV's consequences and how a woman reacts to it cannot be neatly separated out from the other types of violence present in a woman's life. This is equally true for all women globally who exist within violent hierarchies of class, religion, immigration status, age and sexuality, not just gender and ethnicity.[156]

1.4.2.2. Improving Health

Collecting ethnicity data in health research can not only highlight inequalities but also potentially leads to insights into what accounts for differences, providing the possibility of solutions based on effective interventions. This process can improve the health of individuals in different ethnic groups and help overcome health inequalities.[137]

With regards to IPV, it is not enough to know that there are ethnic differences in IPV. To improve health one would then want to disentangle the reasons for these differences in order to respond in a practical way, working towards decreasing IPV. Indeed Bhopal states that

> "The only ethical justification for collecting data by ethnicity and health is health improvement either directly or through research."[134]

Individuals are unlikely to consent to a study which uses their ethnicity data but is not ultimately trying to improve their group's health. Ethical robust research requires consenting participants. The aims for secondary analysis of ethnicity data need to be focussed on health improvement even more so as the use of anonymous data means that individuals may have not consented.

1.4.2.3. Responding to increasing ethnic diversity

Collecting ethnicity data becomes increasingly important as worldwide the numbers of international migrants increases and there is increasing ethnic diversity. In the US 46 million residents speak a different language from their primary care clinicians. Minority groups socially defined by “race” and “ethnicity” will be more than 50% of the population.[157] In England 12.5% of the population (6.4 million residents) is in an ethnic minority. In the UK, 7.6% of the population is from an ethnic minority group, representing an increase of over 44% over the preceding decade.[158] These migrants predominantly come from poor countries (in Africa, Asia, Central and South America) to rich affluent countries. 53% of new immigrants to European countries are women and 50% are women to North American countries.[148] Potential ethnic differences in the experience and nature of IPV are an important aspect of women’s health that should not be ignored.

In the UK, Europe and North America as populations diversify due to the immigration of people from low resource countries, the differences in health between ethnic groups becomes important in order to sustain a fair and just society. In the UK the notion of equitable access to services is firmly embedded in the founding principles of the NHS. An equitable service can only be provided by understanding the differences in patients, including those from different ethnic groups. In the UK the function of strong race relations legislation is to promote equality in a multi-ethnic society.[134] Enshrined within this is that a response is required when health inequalities are manifest. In the US, the “inclusion and difference paradigm” as described by Epstein, represents policy on including diverse groups in medical studies whilst measuring differences across those groups.[140] Having considered the rationale for collecting ethnicity data I now reflect on the dangers of collecting ethnicity data.

1.4.3. Dangers of collecting ethnicity data in health research

The dangers of collecting ethnicity data in research studies include racism, arbitrary classification and inadequate analysis which I now expand upon.

1.4.3.1. Racism

The fundamental potential danger of collecting ethnicity data is that it, as with racial data, can be used to advocate racism. This has been very evident in the past but is still possible in the present, especially when there is no accompanying analysis or interpretation of ethnic differences within studies by researchers (see section 1.4.3.3). This leads the susceptible reader to the conclusion that biological differences account for social inequalities.[140]

Historically racial classifications were thought by scientists to be firmly based on biological facts and were directly used to support racism, a belief in the superiority of a particular race with prejudice based on this. This resulted in antagonism and discrimination towards other human beings (said to belong to different races) with the underlying theory that human abilities are determined by race. There are numerous examples of this, perhaps the most well known being the West African slave trade. This involved approximately 12 million Africans being forcibly removed from their homelands, from 1500 to about 1900, in order to increase the wealth of Europeans. Another example is the extermination of Jews by the Nazis in the 20th century.[159] Perhaps what is less well known is that medicine was not a passive bystander but played a very active role in composing and mitigating these racial hierarchies.[140, 160]

Advocating racism on the basis of scientific fact also persists in current times. For example, the recent widely publicised view of the eminent scientist, Nobel laureate, discoverer of DNA structure, James Watson that black people are less intelligent than white people[161] and the explosion of mainstream media interest about The Bell

Curve.[162] No TV channels gave airplay to other views, for example. that using IQs or “g” scores to rank individuals is a major misuse of science.[160] Scientific racism means the use of science to develop theories and propaganda that draw upon the biology of racial differences. This continues to be active today. Racism is a present day reality throughout the world and not just a historical fact.[163]

1.4.3.2. Arbitrary classification

When classification based on ethnicity data appears in academic journals it appears to have scientific validity. However classification based on ethnicity is subjective, context-specific, purpose-driven and imprecise.[134] It results in idiosyncrasies such as individuals classified as being Hispanic in US studies simultaneously being categorised as white in south American studies. This context specificity of ethnicity is partly a strength which in this example successfully captures cultural positions. However in other more complex examples ethnicity is likely to be too limited to capture all of the nuances and subtleties of cultural position.[164]

Researchers also use census ethnicity categories though these have been developed for administrative reasons. Bhopal stated that ethnicity was not measurable with accuracy or validity[153] and that the classification used for the 1991 UK census was arbitrary. He said that the UK census only worked as the population were willing to answer it, partly as it had been developed with input from ethnic minority organisations.

Guidance on the use of ethnicity in health research published in the British Medical Journal (BMJ) and the Journal of the American Medical Association (JAMA) both emphasise the importance of the terms used to describe ethnicity.[137, 165] The BMJ guidance stressed that the terms used should be descriptive showing how groups were defined with the logic underlying ethnic groupings and their allocation included in the Methods section.[166] Similarly the JAMA guidance highlighted defining categories precisely and being able to state how persons are allocated to these categories.[137] These measures encourage transparency about the arbitrary nature of ethnicity classification systems used by a paper.

1.4.3.3. Inadequate analysis

When ethnicity data appears in papers there is often no analysis by ethnic subgroups. Epstein states that there should be clarity about why a researcher assumes that ethnicity (or race, sex, gender and / or age) are medically meaningful identity attributes in a particular subject area.[140] Often this reasoning is ignored in papers and furthermore what accounts for any revealed ethnic differences is not always addressed. For example, one IPV study, detected a significant difference in the ethnicity of IPV cases compared to controls, along with a number of socio-economic differences.[167] There was no further analysis or discussion about this result. Another IPV study found that African-Americans in Newark, US were significantly more likely to be coded as an IPV visit (odds ratio 1.9) when attending the Emergency Department.[168] Again there was no further discussion about what may account for this result. There are exceptions to these studies, with examples of thoughtful data analysis to account for ethnic differences. For example, a study skilfully investigated mothers' health behaviours by unpacking ethnicity effectively.[169] Thus acculturation indicators (generational status, language spoken at home, length of residency in UK) were also examined when looking at mothers from ethnic minorities. Harmful maternal health behaviours were shown to rise as length of residency in the UK increased. Data is vulnerable to xenophobic interpretation when there is limited analysis. In the first two studies, headline results may convey that some ethnic groups are biologically more violent and in the second that pregnant Asian mothers do not need to be asked about drinking or smoking. As with all other research, researchers looking at ethnic differences need to have focussed research questions which can be addressed by the data collected and data collection needs to include potentially confounding factors that ethnicity may be a proxy for. This is perhaps more important when ethnicity is involved because of the potential for misinterpretation.

In family violence research it has been suggested that eco-cultural factors including economic marginality, salience of religion, social support, domestic and family workload need to be separated out from ethnicity when it is being investigated.[148] These eco-cultural confounding factors allow one to understand what may account for

apparent ethnic differences that may be seen in family violence. They are consistent with the contributing factors put forward by the Ecological Model of IPV (see section 1.1.). In IPV identification, ethnicity may also be confounded by socioeconomic status. For example, a higher IPV prevalence in a particular ethnic group may be related to the group's lower socioeconomic status. This would potentially increase the PPV of index questions in this ethnic group which could actually be due to a difference in socioeconomic status rather than an ethnic difference per se.

The British Medical Journal[165] and the Journal of the American Medical Association[137] guidance on the use of ethnicity in health research both include that ethnicity may be confounded by socioeconomic status. The need to adjust either for social class or socioeconomic status has been depicted as "a necessary first step" in investigating ethnic differences.[170] Yet socioeconomic status as a confounder is often neglected in comparisons between ethnic groups[153] reflecting inadequate analysis.

1.4.4. Five criteria to assess the use of ethnicity data by papers

I used published guidance on the use of ethnicity in health research from the British Medical Journal,[165, 166] the Journal of the American Medical Association[137] and more specific guidance related to IPV[148] to generate five criteria which indicate how effectively papers handle ethnicity data. This is one way of appraising the quality of a paper. My five criteria encompass what I believe is the minimum that investigators should address if they have chosen to collect ethnicity data in their research. I produced the five criteria by examining the common themes arising from this published guidance whilst endeavouring to isolate the most important issues. See Box 1, on page 85, for my five criteria.

Box 1:

Five criterion checklist (DECSS) for quality appraisal of the use of ethnicity data

1. **D**: Is ethnicity **described**?
2. **E**: What are the terms used to describe **ethnicity**?
3. **C:** Is the **classification** system using ethnicity justified?
4. **S**: Is ethnicity **self**-assigned?
5. **S**: If the study analyses differences in ethnic groups are **socio-economic** factors considered or controlled for?

These five criteria take into account the rationale for collecting ethnicity data (see section 1.4.2.) whilst trying to minimise the dangers of collecting ethnicity data (see section 1.4.3.). Consequently these criteria scrutinise what and how ethnicity terms are used in papers. The importance of justifying the classification used (for example by having an underlying hypothesis) and self-assignment is emphasised. If ethnic differences are being investigated, the need to consider socioeconomic status as a potential confounding factor is highlighted. For further details about these criteria see Method, section 2.2.5.

Having described the background to the meaning of ethnicity, I now consider IPV research that has used ethnicity data. This is followed by a description of the clinical problem which I address with my research questions. My principal research question and a related second research question are presented, followed by my study aims and a thesis outline.

1.5. IPV Research and Ethnicity

IPV prevalence studies from around the world show that IPV is common in many different ethnic groups.[6, 8, 123, 171-176] Studies from the US and UK have made within-study ethnic comparisons. These generally have not found differences between ethnic groups with respect to IPV prevalence, pattern or severity in abused women.[171, 172, 174, 175, 177] Exceptions to this include two large studies, (one with a study population of 16,000) which found apparent ethnic differences in IPV prevalence in the US.[1, 178] On analysis these ethnic differences were accounted for by lower income[1] and lower education levels.[178] Other studies which have showed ethnic differences in IPV in the US have not analysed or discussed this variation any further.[167, 179] Campbell and colleagues[167] using a case control study showed that the percentage of IPV cases that were African-American was greater than white. Dearwater and colleagues[179] using logistic regression analysis showed that African-American ethnicity was an independent risk factor for lifetime emotional or physical abuse.

Factors which have been suggested as accounting for variation in IPV within the same ethnic group are lower income in black women,[180] perceived racial discrimination in black women in New York[181] and immigrant related factors (social isolation, lack of awareness about IPV services, immigration policies preventing women on spousal visas from working and emotional abuse from in-laws) in south Asian female immigrants in Greater Boston.[182-184] The effects of acculturation (generational status reflected by country of birth and language) were not consistent in either the same or different ethnic groups.[173, 182, 185-188]

The large multi-country WHO study was powerful and found globally that there were wide differences in prevalence of IPV, patterns of IPV and attitudes towards IPV both between and within countries. This landmark study directly measured the extent of IPV experienced by 24,000 women, from 15 sites in 10 countries. It used cross-sectional population-based household surveys which allowed comparison and analysis across different settings, ensuring that variations mostly signified real differences – unlike earlier work.[6] The countries studied included Bangladesh (with urban and

rural study sites), three African nations and Serbia – all of which have migration to urban first world countries. There was awareness that when IPV is measured in these very different cultures that disclosure would always be affected by cultural beliefs and biases. The standardised methodology, measuring IPV cross-culturally and using conservative definitions of violence helped to ensure that real differences were measured.

The WHO study found that the prevalence of lifetime physical violence and sexual violence by an intimate partner among ever-partnered women varied from 15% in Japan to 71% in Ethiopia. This wide variation existed not only from country to country but also between urban and rural provincial settings within countries. For example the percentage of women affected from rural Bangladesh was 62% versus 54% from urban Bangladesh. In most settings, sexual IPV was less frequent than physical IPV except in provincial Bangladesh, Ethiopia and urban Thailand where it was more frequent. The differences between settings were not accounted for by socioeconomic factors alone though age, marital status and education level did cause some variation in IPV prevalence.

The wide contrasts seen in women's attitudes towards IPV suggested that there were cultural differences between the study populations sampled. These were most marked between the urban, industrialized settings and the rural, traditional ones. In rural Bangladesh 80% of women agreed that wife beating was justified for certain reasons; with ~15% believing that a woman did not have the right to refuse to have sex with her partner under any circumstances, even if he mistreated her. Women in poorer countries were more likely to think that violence was justified, with the highest rates being in more rural traditional communities where the problem remained largely hidden. Half of the women surveyed had never spoken of their situation to anyone. Some said they did not report the violence because they considered it normal. Some even said their husbands were justified in beating them, illustrating an impact of male domination.

Another multi-country study has also demonstrated a wide variation in women's attitudes towards violence.[189] In India, 70% of women believed that wife beating was justified for at least one reason whilst in the Dominican Republic this figure was

11%. I think that it is possible that some of the attitudes seen in poorer developing countries persist in diaspora populations who have moved to richer industrialised nations. This may affect how individuals respond to questions asking them about IPV as well as how they define IPV.

Strong beliefs about the importance of male domination along with the woman's responsibility to keep the family together and the centrality of the family have certainly also been expressed by African-American women,[190] Hispanic Mexican-American women, Anglo-American women,[191] Japanese-American women in the US[192] as well as Japanese women in Japan,[193] south Asian women [194] and Jewish Israeli women.[195] These beliefs about male domination have been presented in all these individual studies bar one as specific cultural beliefs exclusive to that particular ethnic group. I think that this is an example of the "cultural deviant perspective" in which there is an overemphasis on the role of cultural values in propagating IPV in different ethnic groups.[196] Others have also pointed out that culture should not be held to account for all the variation seen in patients.[152] Almost universal states of male domination and family centrality should not be attributed only to certain cultures.

I think that these studies collectively support the universality of the gender power imbalance rather than gender power imbalance being a cultural belief exclusive to a group as suggested by the authors of many of these individual papers. Recalling the universality of the imbalance in gender power helps to guard against cross-cultural hypocrisy over gender practices.[197] It is important that gender based violence in majority communities is not hidden away[128] whilst being exposed in minority groups.

1.5.1. Research questions

The authors of the two studies examining Japanese-American women and south Asian women in Boston, proposed that there were ethnic differences in attitudes towards IPV.[183, 188] Similarly Torres and colleagues[191] showed that despite there being no significant difference in the severity or frequency of wife abuse between Mexican-American and Anglo-American women that there were differences in what was perceived to be abusive and the likelihood of reporting abuse. Anglo-Americans were more likely to label specific behaviours (for example being punched, slapped or pushed) as abuse than Mexican-American women. Or put another way Mexican-American women if asked whether they were being abused were less likely to say yes even if they were experiencing some of the above actions from their partner. This clearly shows that the precise wording of questions trying to identify IPV in different ethnic groups in a health setting could have a drastic effect on the identification rate of IPV. Close attention needs to be paid to the content validation of questions trying to identify IPV in health care settings. So in this example, it would be better to ask these women about specific behaviours or impacts of IPV rather than asking "Have you been abused?"

There is a large volume of literature describing questions that can be used to identify IPV in both clinical and research settings. Whether any studies address how questions trying to identify IPV function in individual ethnic groups is unclear. Ethnic differences in the ability of questions to identify IPV are alluded to in the literature but it is uncertain whether this has been methodically investigated. Hence it has been reported that the utility of validated tools to detect abuse in diverse populations is unknown,[198] with self report surveys mostly validated among white populations.[196] Sorenson concurs that survey instruments for IPV which have been developed and used with "Anglos" have then been directly used in other ethnic groups [199] without assessing whether it is valid to do so. Therefore the clinical problem for health care professionals and researchers of IPV is are the questions that they use to identify IPV valid in women from different ethnic groups. I address this problem with my two research questions.

My principal research question is:

- What is the evidence for the validity of questions trying to identify IPV in specific ethnic groups?

Related to this is my second research question:

- Does the evidence for the validity of questions trying to identify IPV vary between different ethnic groups?

1.5.2. Aims and objectives

Intimate partner violence (IPV), including physical, sexual and emotional violence, causes short and long term ill-health.[15] Brief questions that can identify women from different populations experiencing IPV who present in clinical settings are a pre-requisite for an appropriate response from health services to this substantial public health problem.[200]

My principle research aim is to examine the evidence for the validity of questions trying to identify IPV in different ethnic groups. My second aim is to determine whether these questions' validity varies between ethnic groups.

My research objectives are to firstly systematically search the literature for index questions for the identification of IPV in different ethnic groups and assess their evidence of validity. My second objective is to analyse the data from a cross-sectional survey of four questions (HARK) identifying IPV in a primary care study population.[200] This is to generate diagnostic indices for identifying IPV in three ethnic groups for the four HARK questions; and then to generate diagnostic indices for the dimensions of IPV (physical and emotional abuse) in three ethnic groups for individual HARK questions.

1.5.3. Outline of the thesis

Chapter 1 has presented the case for IPV being a priority that should be addressed by primary care. This background chapter has also covered the central theoretical concepts that underpin my thesis. These include the background to measuring validity and the meaning of ethnicity. There has been consideration about the different methods to measure the validity of questions trying to identify IPV. There has also been reflection on the rationale as well as the dangers of collecting ethnicity data in health research. IPV research that has used ethnicity data has been presented. This subsequently has led to the articulation of my principal research question and my related secondary research question.

Chapter 2 describes the methods used to answer my two research questions. This includes the systematic review used to identify relevant research papers as well as my secondary analysis of data generated by a cross sectional survey.

Chapter 3 presents the results of the systematic review and my secondary data analysis. The systematic review's results are presented in a series of tables with accompanying narrative results. The quality of the methodology is appraised using QUADAS, a 14 item tool for Quality Assessment of Diagnostic Accuracy Studies. The quality of the use of ethnicity data is appraised using five criteria generated from existing published guidance. The narrative account complements the tabulated results by justifying decisions made about the QUADAS criteria and containing further information on how ethnicity data were used by primary studies. The results of my secondary data analysis are presented predominantly in tables with complementary receiver operator characteristic curves and a brief commentary.

Chapter 4 first summarises the answers to my research questions and then considers why these results are important by considering them in the context of other reviews and clinical practice. Following this there is evaluation of the quality appraisal of my methodology and the quality appraisal of how ethnicity data is used in the systematic review studies and my secondary data analysis. The limitations of QUADAS as a

quality appraisal tool are discussed. The overall strengths and limitations of my thesis are examined.

Chapter 5 presents my conclusions, followed by recommendations for future research in this field.

Thus in this background chapter, I have justified the case for IPV to be a priority in primary care, considered how to measure validity, reflected on the meaning of ethnicity and presented my research questions, aims, objectives and thesis outline. In my next chapter I will describe the methods used in my research.

Chapter 2: Methods

2.1. Overview

I used a systematic review and a cross sectional survey to evaluate the items used in tools to identify IPV in specific ethnic groups in order to answer my principal research question and my related secondary research question:

- What is the evidence for the validity of questions trying to identify IPV in specific ethnic groups?
- Does the evidence for the validity of questions trying to identify IPV vary between different ethnic groups?

Both of these research questions were addressed using both of the methods – the systematic review and my secondary data analysis of the cross sectional survey. The cross-sectional survey originally estimated the diagnostic accuracy of HARK (four questions trying to identify IPV) in a study population of varied ethnicity.[200]

In this chapter I describe both of these two methods. The use of two different approaches to answer the same two research questions is a deliberate strategy that I have employed to try to improve the quality of the answers generated. The research answers will be more robust if they embody converging results from independent methods. Using more than one method is thought to lead to a greater understanding of a subject than if a single method is used whilst "challenging conventional thinking…offering multidimensional insights."[201]

2.2. Systematic review

The aim of the systematic review was to find, evaluate and synthesise research looking at questions used to identify IPV and to consider this research's relevance to identifying IPV in different ethnic groups.

There have been many studies validating questions to identify IPV which often collect ethnicity data. It is uncertain whether ethnic differences in the ability of questions to identify IPV have been examined. Systematically reviewing these research papers may help to establish whether evidence of validity exists for questions trying to identify IPV in specific ethnic groups, the quality of this validity evidence and whether this evidence varies between different ethnic groups.

2.2.1. Data sources and search strategy

I searched nine electronic databases for relevant papers which tried to assess the validity of questions to identify IPV. The electronic databases were all searched from their individual respective start dates until 31st of December 2007.

The nine electronic databases: Cochrane Collaboration central register (CENTRAL/CCTR), Medline, Cumulative Index to Nursing and Allied Health Literature (CINAHL), British Nursing Index (BNID), Embase, National Research Register (NRR), Health Management Information Consortium (HMIC), Midwives Information and Resource Service (MIDIRS), NHS Economic Evaluation and Database of Abstracts of Reviews for Effectiveness (DARE).

I found eligible studies using pre-defined search strategies which had previously been used in a related systematic review.[7] These predefined search strategies used a mixture of content terms and test types. See Appendix C for search strings. Primary studies describing validation of questions trying to identify IPV were sought. Backward and forward citation tracking were used and examination of papers in

references of included papers to identify further studies. The first or corresponding authors of included papers were contacted to try to identify further relevant studies. The authors of the four papers which were found in the last updated search covering the period from December 2006 to December 2007 were not contacted due to time limitations apparent at that stage. International researchers of IPV were contacted using partner violence organisations and research networks in the UK, Europe, US and Australia in order to identify relevant papers. There was no hand searching of journals.

2.2.2. Study selection

The inclusion criteria were:

- The study participants had to be aged over 15 years.
- The study design had to include validation of the items in the IPV identification tool, compared to another tool.
- The comparator tool was either a standard reference criterion or other test intended to measure a construct which was similar or related to IPV. There was no limit on the number of questions that the comparator tool contained.
- The index IPV tool being evaluated had to contain less then eleven questions and so be short enough to be potentially used in routine ten minute primary care consultations as part of the clinical history taken by clinicians.
- The outcome measures needed to include either indices of diagnostic accuracy (i.e. sensitivity, specificity, PPV, NPV, LR, PTO or ROC curves) or correlation coefficients (representing the relationship between index questions and comparator tools) or reliability measures of the index questions.
- The study setting could be in or outside a health care setting. There were no restrictions on the geographical or national setting.
- Studies published in peer-reviewed journals or in published books.

The exclusion criteria were:

- Studies in which the sole participants were male survivors of partner abuse of any age.

- Studies which involved the survivors of abuse committed by other family members (such as in-laws), studies reporting joint treatments, such as couple or family therapy (even if the therapy was administered separately to women) and community or societal interventions conducted with the aim of increasing public awareness of the problem of partner abuse.
- Papers published in non-European languages.
- Studies that used non standardised interviews as a comparator with no known sensitivity, specificity or reliability at identifying IPV.

2.2.3. Data extraction

All the eligible papers were read with relevant data recorded on to electronic data collection forms, (see Appendix D). Summary data were entered into summary tables (see Results, tables 1 to 6, pages 110 to 133).

2.2.4. Analysis of primary data extracted

For each set of index IPV questions being evaluated, evidence of validity was collected including any diagnostic accuracy indices, correlation coefficients and reliability measures. I also examined the validity evidence based on response processes of the index questions and the content validation of the index questions. This was followed by synthesis and interpretation of the data collected.

I considered combining the results from studies about the same index questions in specific ethnic groups by pooling data if primary studies contained the same outcome measures for the same index questions. However this was not possible as primary studies did not contain the same index questions for specific ethnic groups with the same outcome measures.

2.2.5. Quality appraisal

The quality of studies was appraised using the 14-item Quality Assessment of Diagnostic Accuracy Studies (QUADAS) [202-204] and by assessing the appropriateness of authors' use of ethnicity data, using published guidance on the use of ethnicity in health research.[137, 148, 165, 166]

The QUADAS tool has been specifically devised to be used in systematic reviews to assess the quality of primary studies of diagnostic accuracy.[204] Hence it was thought to be appropriate to use to appraise the studies in this systematic review of questions identifying IPV in specific ethnic groups. I was unable to find any published tools to assess the quality of studies which used the validation paradigm. The validation paradigm presumes that there is no existing gold standard which can identify IPV (see section 1.3.3.).

QUADAS rates studies for bias (8 items), variability (1 item), and reporting (5 items). This includes examining patient spectrum, selection criteria, reference standard, partial verification, differential verification, incorporation, test execution, blind analysis, interpretation, indeterminate results and study withdrawals. See table 5 for the complete QUADAS criteria. The QUADAS outputs are descriptive results relating to the potential sources of bias. There is no scoring system and no generation of a single score.[205] I used QUADAS to assess within-study bias, looking at the level of methodological quality of each primary study.

How studies used ethnicity data (i.e. what ethnicity terms were used, how these ethnicity terms were used and whether confounding was considered) was assessed by applying criteria which I devised but which originated from published guidance on the use of ethnicity in health research.[137, 165, 166, Malley-Morrison, 2007 #440] See section 1.4.4. My criteria are listed on page 98.

Five criterion checklist (DECSS) for quality appraisal of the use of ethnicity data

1. **D**: Is ethnicity **described**?
2. **E**: What are the terms used to describe **ethnicity**?
3. **C:** Is the **classification** system using ethnicity justified?
4. **S**: Is ethnicity **self**-assigned?
5. **S**: If the study analyses differences in ethnic groups are **socio-economic** factors considered or controlled for?

The first two criteria try to achieve clarity on whether ethnicity is described and how it is described. They are checking whether papers are being clear about the arbitrary nature of ethnicity classification. The third criterion on justifying the classification system used (for example by having an underlying hypothesis) is to assess whether researchers have been clear about why ethnicity data has been collected. If ethnicity is thought to be an important factor, there should be explanation about why this is the case. Alternatively the justification for the ethnicity classification system used may be to assess whether the study population is representative of the actual population. The fourth criterion highlights the importance of self-assignment in order for ethnic classifications to have any validity. The final criterion is to ensure that if ethnic differences are considered that there is adequate basic analysis of ethnicity data. Studies that characterise the ethnicity of participants but not their socioeconomic status are at risk of confounding ethnicity with socioeconomic status.

2.3. Secondary data analysis

The aim of my secondary data analysis of a cross-sectional survey was to investigate how four questions used to identify IPV (HARK), performed in different ethnic groups. The cross sectional survey sampled women waiting to be seen by a doctor or nurse, sitting in general practice waiting rooms. This survey's original study aim was to validate the use of HARK in primary care. Ethnicity data was collected to see if the study population was representative of the local population.[200] For a copy of this published paper, see Appendix B. The potential value of pre-existing data has been recognised.[206] My exploratory secondary data analysis of the HARK study allows the diagnostic accuracy of the four HARK questions to be directly calculated in specific ethnic groups and then compared to see if there are any differences which could potentially be clinically important, for example if a particular question did not identify IPV in a specific ethnic group.

Hence my secondary data analysis investigated:

- the four HARK questions' ability to identify IPV in the different ethnic groups and whether this varied in the different ethnic groups
- the individual HARK questions' ability to identify IPV in the different ethnic groups and whether this varied in the different ethnic groups
- the individual HARK questions' ability to identify specific dimensions of IPV (emotional IPV and physical IPV as defined by the CAS) in the different ethnic groups and whether this varied in the different ethnic groups

The HARK study data were collected from May 2003 to October 2003. The HARK study population was sampled from all women sitting in selected GP waiting rooms over this time period. The fifty-one general practices in Newham, a multi-ethnic inner city area of London, were stratified according to the number of doctors and the proportion of south Asian names on the practice register.[207] Equal numbers of practices were selected from each stratification group using a randomisation procedure within a statistical software package for social sciences (SPSS version 12). This was in an attempt to ensure that the practice population reflected the local area population.

Out of all the women sitting in the GP waiting rooms, seven hundred and thirty seven women did not meet the inclusion criteria. Fourteen women were not approached because there were too many women in the waiting room for all the women to be approached. Two hundred and three women said that they would participate in the study but then did not come back following their clinical consultation. One hundred and eighty six women declined participation in the waiting room, only knowing that the survey was about women's health. Eleven women declined consent in the private room, knowing that the study was about IPV. In total 232 women agreed to participate and completed the survey. The response rate of 54% (232 / (232 + 186 + 11) was adjusted for the women who did not come back following their clinical consultations (see figure 6, page 101). The unadjusted response rate was 36%. This included in the denominator the women who were not approached and women who said that they would participate in the study but then did not come back following their clinical consultation.

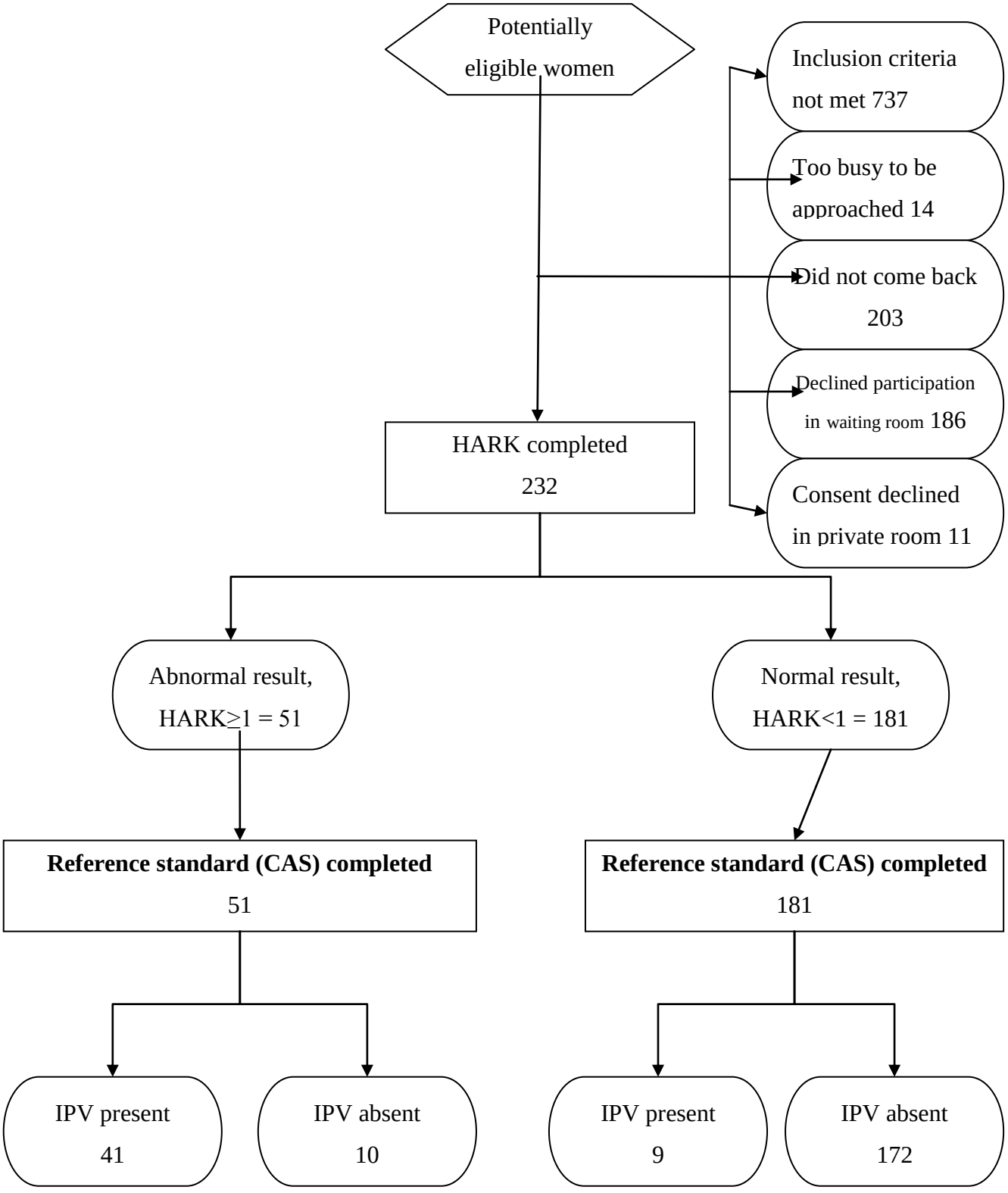


Figure 6: Flow diagram of recruitment of participants to the study

The 232 women included in the current analyses were more than 17 years old and in the last year had been in an intimate relationship. The research protocol for the study was approved by The East London and City ethics committee. Further details of the method for this cross sectional survey, including the data collection and the data analysis are described elsewhere.[200] This published paper is reprinted in Appendix B. It should be noted that more in depth coverage of how the HARK questions were developed (for example details of the pilot study) are contained in my MSc thesis.[95] These details have been deliberately omitted from my current work, avoiding repetition.

2.3.1. Sample size

The HARK study was originally powered so that the sample size used demonstrated an acceptable level of sensitivity for the four HARK questions.[200] This further secondary exploratory analysis was underpowered to detect statistically significant ethnic differences in the diagnostic accuracy of HARK. I amalgamated data into larger groups in an attempt to increase the power to make comparisons between groups. Black British, African and Caribbean were combined to form an African-Caribbean group. Indian, Pakistani, Bangladeshi and Sri Lankan data were combined to form a south Asian group. White British, white Irish and white other were combined to form a white group. My logic underlying these ethnic groupings was that women within each amalgamated group (African-Caribbean, south Asian and white) may share cultural beliefs (about gender roles, expectations about family roles, what is considered to be abusive, willingness to disclose abuse and reasons for disclosing abuse – see Background, section 1.4.1.3.) affecting how they respond to questions asking about IPV.

A power calculation showed that with the sample sizes of my three amalgamated groups my analysis had a 80% power to detect a 17% difference in PPV between the African-Caribbean and white groups (79% to 96%); and a 80% power to detect a 19% difference in PPV between the south Asian and white groups (79% to 98%).

My use of 95% confidence intervals allows the level of precision of the results to be defined. Inspecting the limits of the 95% confidence limits of my results enables the planning of a future larger authoritative study into the ethnic differences in the validity of questions identifying IPV.

2.3.2. Analysis

My secondary data analysis examined the combined four HARK questions' ability as well as the individual HARK questions' ability to identify IPV and its dimensions, in the different ethnic groups and then whether this varied in the different ethnic groups. Statistical analyses of my data, including generating ROC curves, were conducted using STATA, a statistical software package.

2.3.2.1. HARK's ability to identify IPV in different ethnic groups

The rate of current IPV with 95% confidence intervals, within the last twelve months was calculated for the CAS (using the cut off score of ≥ 3) within the three main aggregated ethnic groups: i. African-Caribbean (black British, African or Caribbean) ii. south Asian (Indian, Pakistani, Bangladeshi or Sri Lankan) and iii. white (white British, white Irish or white other). For each of these three main ethnic groups, the rates of IPV within the last twelve months were also calculated with 95% confidence intervals for the HARK, at different cut off scores (for example, HARK cut off score ≥ 2, means a HARK score of either 2, 3 or 4).

Each woman was identified as being positive or negative for IPV for each HARK cut off score and for the CAS cut off score of ≥3. I then calculated within each ethnic group HARK's sensitivity, specificity, PPV, NPV, LR (all with 95% confidence intervals) and post-test odds (= pre-test odds x LR) at different HARK cut off scores.[91] The change from the pre- to post-test probability of IPV that occurred using different HARK cut off scores was then calculated.

For each ethnic group, a ROC curve was constructed by plotting the sensitivity of each different HARK cut off against the false positive rate (= 100 – specificity) at the different HARK cut offs. This was used to determine an optimal cut off HARK score and the instrument's overall sensitivity and specificity in each ethnic group. In each ethnic group at the optimal HARK score, the estimates for each of the diagnostic indices and their corresponding 95% confidence intervals were inspected to check whether any were not overlapping, representing statistically significant differences.

2.3.2.2. Each individual HARK question's ability to identify IPV in different ethnic groups

For each of the three main ethnic groups, each woman was identified as being positive or negative for IPV, according to whether each individual HARK question was positive or negative, as well as for the CAS cut off score of ≥3.

I then calculated within each ethnic group each individual HARK question's sensitivity, specificity, PPV, NPV, LR (all with 95% confidence intervals) and post-test odds (= pre-test odds x LR) for identifying IPV. The difference between the pre-test probability and the post-test probability of IPV was also calculated. This allowed us to examine for example, whether being humiliated ("H") was more predictive of IPV in the African-Caribbean group as opposed to the south Asian group. For each individual HARK question, the estimates for the diagnostic indices and their corresponding 95% confidence intervals were inspected to check whether any were not overlapping, representing statistically significant differences.

2.3.2.3. Each individual HARK question's ability to identify dimensions of IPV (emotional and physical IPV) in different ethnic groups

Furthermore, for each of the three main ethnic groups, each woman was identified as being positive or negative for emotional IPV (according to whether emotional IPV

was present, as defined by whether the score for the CAS emotional abuse score was ≥ 3) and physical IPV (as defined by whether the score for the CAS physical abuse score was ≥ 1).

I then calculated within each ethnic group each individual HARK question's sensitivity, specificity, PPV, NPV, LR (with 95% confidence intervals) and post-test odds (= pre-test odds x LR) for identifying emotional and physical IPV, two dimensions of IPV. Additionally the difference between the pre-test probabilities and the post-test probabilities of emotional and physical IPV when identified by individual HARK questions were calculated. This allowed us to examine for example, whether being humiliated ("H") or being afraid ("A") was more predictive of emotional IPV in the African-Caribbean group as opposed to the south Asian group. For each individual HARK question used to identify either emotional or physical IPV, the estimates for the diagnostic indices and their corresponding 95% confidence intervals were inspected to check whether any were not overlapping, representing statistically significant differences.

Having described the methods used in my research, in the next chapter I will present my results.

Chapter 3: Results

3.1. Overview

This results chapter will present the results of the systematic review followed by the results of my secondary data analysis of a cross-sectional survey. The cross-sectional survey used four questions (HARK) to identify IPV in a primary care sample of women. The systematic review's results have been presented in a series of tables with an accompanying narration. The secondary data analysis results are presented in tables of diagnostic indices with 95% confidence intervals and figures of receiver operator characteristic curves with a brief commentary.

Both the results of the systematic review and the secondary data analysis are organised in order to try to answer my principal research question (what is the evidence for the validity of questions trying to identify IPV in specific ethnic groups?) and my related secondary research question (does the evidence for the validity of questions trying to identify IPV vary between different ethnic groups?).

3.2. Systematic review results

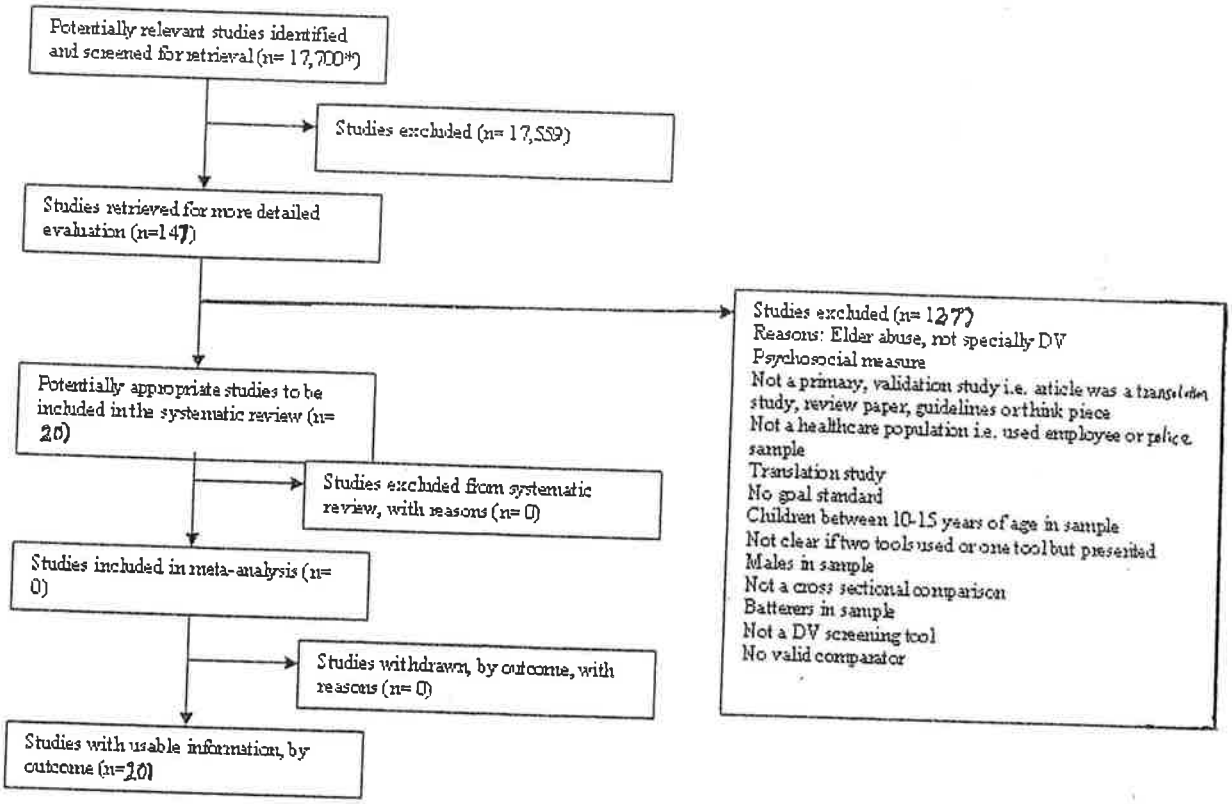
The chief findings of the systematic review are presented in a series of tables with accompanying narrative results. Table 1 shows a summary of the results of the twenty studies in the systematic review. Table 2 contains further details about the characteristics of each study (for example study design and study population description). Table 3 shows the correlation measures used for criterion correlation validity and convergent validity with their interpretation by papers. Table 4 shows the correlation measures used for internal consistency reliability with their interpretation by papers. Table 5 summarises the QUADAS quality items for each study. Table 6 summarises the criteria assessing the use of ethnicity data for each study. Following these tables, the narrative results include a brief overview for each study. The narrative results provide material which is additional to the tabular results. Thus there is explanation about the rationale for potentially contentious judgements made with respect to the QUADAS criteria (see table 5) and information, in addition to table 6, about how ethnicity data was used by studies. The tables should be read alongside the narrative results. To accommodate this, the overall summary of the results shown in table 1 is presented in a laminated version and placed in the front sleeve of this thesis so that it can be easily referred to when reading the narrative results.

This systematic review identified 20 validation studies reported in 20 papers assessing 17 sets of questions trying to identify IPV.[198, 200, 208-225] These sets contained a total of 76 questions trying to identify IPV. These studies involved validation in 11,648 participants. The publication dates ranged from 1992 to 2007. A flow diagram showing the numbers of studies retrieved at each stage of the systematic review is shown in figure 7, on page 109. In total 147 full papers were read. 127 studies were excluded after applying the exclusion criteria. 20 studies were included which fulfilled the inclusion criteria.

Out of the 20 studies, 11 studies explicitly used the classical diagnostic accuracy paradigm, generating diagnostic accuracy data.[198, 200, 208-216] Five of these studies contained only diagnostic accuracy data.[200, 208-211] Eight studies used methods from the validation paradigm as well as the diagnostic accuracy

method.[198, 212-218] Two of these studies reported the figures which allowed the computation of diagnostic accuracy results.[217, 218]

Out of the 20 studies, 14 studies used the validation paradigm, including the eight above that also contained diagnostic accuracy data.[198, 212-224] Eight studies out of the 20 contained eleven estimates of internal consistency reliability for eight sets of questions trying to identify IPV.[198, 214-216, 219-222] Six studies focussed solely on the validation paradigm.[219-224] Four of them used a variety of methods, including convergent validity, known group comparisons and internal consistency reliability.[219-222] Two used only a convergent validity method.[223, 224] The remaining one study used neither of the research paradigms.[225]



* This figure represents the total number of abstracts identified by search strings.

Figure 7: Flow diagram of studies retrieved for systematic review

Table 1: Summary results of 20 studies in systematic review

Study	Index Questions Evaluated	Comparator Tool	Measures of Validity									
			Diagnostic accuracy indices							Criterion or Convergent validity (correlation or %)	Internal consistency reliability	Response processes validity	Content validation
			Prevalence of IPV (%)	Sen (%)	Spec (%)	PPV (%)	NPV (%)	LR+	PPV – prevalence				
Five diagnostic accuracy studies													
Peralta & Fleming (2003)	1 question "Do you feel safe at home?"	6 question Modified Conflict Tactics Scale (6item mCTS)	6 item mCTS for any type of violence (period 90 days) 44	9	96	63	57	3	19	Not stated	Not stated	Good – simple scoring system	-
Paranjape, Rask, & Liebschutz (2006)	3 question Slapped, Threatened or Thrown (STaT)	30 question Index of Spouse Abuse (ISA)	ISA Most recent relationship: 33 Current IPV 15	If STaT ≥ 1 (set criterion for a positive) vs ISA						Not stated	Not stated	Good – simple scoring system	Decreased – no sexual IPV question
				95	37	42	94		9				
				If STaT ≥ 2 vs ISA									
				85	54	48	88		15				
				If STaT = 3 vs ISA									
				62	66	47	78		14				
Feldhaus, Kozioi-McLain, Amsbury, Norton, Lowenstein & Abbott (1997)	3 question Partner Violence screen (PVS)	30 question ISA 16 questions: 16 item mCTS	ISA: 24 16 item mCTS: 27	PVS vs ISA						Not stated	Not stated	-	Decreased – no sexual IPV question
				65	80	51	88	3	27				
				PVS vs 16 item mCTS									
				71	84	63	89	5	36				
MacMillan, Wathen, Jamieson, Boyle, McNutt, Woster, Lent & Webb (2006)	3 question PVS 8 question Woman Abuse Screening Tool (WAST)	30 question Composite Abuse Scale (CAS)	CAS ≥ 7: 10	PVS vs CAS						Not stated	Not stated	- For WAST: Decreased – uses term abuse and likert type scale	For PVS: Decreased – no sexual IPV question For WAST: see Brown 2000 study
				49	94	47	94		37				
				WAST vs CAS									
				47	96	55	94		45				

Sohal, Eldridge & Feder (2007)	4 question Humiliation, Afraid, Rape, Kick(HARK)	30 question Composite Abuse Scale (CAS)	CAS ≥ 3: 23	HARK ≥ 1((optimal cut off, maximising true positives whilst minimising false positives) vs CAS						Not stated	Not stated	Good, simple scoring system	Good – discrete questions on physical, sexual & emotional IPV
				81	95	83	94	16	60				
Eight studies using diagnostic accuracy and validation paradigm methods													
Tiwari, Fong, Chan, Leung, Parker & Ho (2007)	3 individual questions derived from the Chinese Abuse Assessment Screen (AAS)	3 subscales of physical, emotional & sexual IPV of 39 question revised Chinese CTS (CTS2)	CTS2 for: Physical IPV: 12 Emotional IPV: 57 Sexual IPV: 9	For physical IPV: 1 question AAS vs CTS2 physical violence subscale						Kappa, 1 question AASs vs CTS2 subscales: Physical IPV 0.56 Emotional IPV 0.52 Sexual IPV 0.47	Not stated		Verified by IPV researchers - 3 nurses, 2 drs, 1 clinical psychologist & 1 social worker
				45	99	87	93	51	75				
				For emotional IPV: 1 question AAS vs CTS2 emotional aggression subscale									
				66	89	89	66	6	32				
				For sexual IPV: 1 question AAS vs CTS2 sexual coercion subscale									
				36	99	80	94	43	71				
Reichenheim & Moraes (2003)	1 question: Portuguese AAS's anchor question on physical IPV during pregnancy	12 question physical violence scale of modified revised Portuguese CTS2 (12 item mCTS2)	12 item mCTS2 for Minor physical IPV 18% Major physical IPV 8% Overall – 19%	For minor physical IPV: 1 question AAS vs 5 items mCTS2						Point biserial correlation, 1 question AAS vs 12 item mCTS2: 0.68	Not stated		Decreased – single question does not consider sexual or emotional IPV
				32	99	88	13	32	70				
				For severe physical IPV: 1 question AAS vs 7 items mCTS2									
				61	98	70	0	28	62				
				For both minor & severe physical IPV: 1 question AAS vs 12 items mCTS2									
				32	99	90	14	40	71				
Ernst, Weiss, Cham, Hall & Nick (2004)	4 question Ongoing Violence Assessment Tool (OVAT)	30 question Index of Spouse Abuse (ISA)	ISA: 20%	OVAT vs ISA						Kappa, OVAT vs ISA: 0.58	Cronbach's alpha 0.6 Inter-item correlation 0.38	Decreased – uses a Likert scale for 1 question	Decreased – does not explicitly consider sexual IPV
				86	83	56	96	5	36				
Chen, Rovi, Vega, Jacobs & Johnson (2005)	4 question English HITS (E.HITS) 4 question Spanish HITS (S.HITS)	11 question ISA- Physical (ISA-P), 8 question WAST; English & Spanish forms	ISA-P: 5% WAST: 10%	English HITS > 10.5 vs English ISA-P						E.HITS vs E.ISA-P 0.76	Cronbach's alpha for: E.HITS 0.76 S.HITS 0.61	Decreased – uses a Likert scale	Decreased - does not explicitly consider sexual IPV
				86	99	86	99	91	~81	E.HITS vs E.WAST 0.75			
				Spanish HITS > 5.5 vs Spanish WAST						S.HITS vs S.ISA-P 0.81			
				100	86	45	100	7	~35	S.HITS vs S.WAST 0.78			

Sherin, Sinacore, Li, Zitter & Shakil (1998)	4 question English HITS	15 question modified CTS (15 item mCTS)		Phase 2, in known group comparison: HITS > 10.5* vs 15 item mCTS						Phase 1: HITS vs 15 item mCTS 0.85	Phase 1: Cronbach's alpha 0.80	Decreased – as above	Decreased – as above**
				96	91	87	97	-	-				
Zink, Levin, Putnam & Beckstrom (2007)	5 non-graphic questions	39 question revised CTS (CTS2)	CTS2: 11%	Question 1. ("how do you and your partner work out arguments") vs CTS2						Not stated	Cronbach's alpha for the 5 questions 0.46	Decreased – uses a Likert type scale	Decreased – does not explicitly ask about physical or sexual IPV (as non-graphic)
				25	98	58	91		47				
				Questions 1. 3. & 4. vs CTS2									
				45	95	51	93		40				
Bonomi, Thompson, Anderson, Rivara, Holt, Carrell & Martin (2006)	5 question Behavioural Risk factor Surveillance Survey (BRFSS)	10 question Women's experience with battering scale (WEB)	WEB (IPV of any type, in most recent relationship): 7%	BRFSS vs WEB:						BRFSS+/WEB+ 5% BRFSS+/WEB- 9% BRFSS-/WEB+ 2% BRFSS-/WEB- 83%	Not stated	Good, simple scoring system	Satisfactory – items on physical, sexual and emotional IPV
				72	90	34	98	7	27				
Coker, Pope, Smith, Sanderson & Hussey (2001)	10 question Women's Experience with Battering Scale (WEB)	15 question modified Index of Spouse Abuse-Physical (15 item mISA-P)	15 item mISA-P 11%	WEB vs 15 item mISA-P:						WEB+/mISA-P+ 9% WEB+/mISA-P- 8% WEB-/mISA-P+ 1% WEB-/mISA-P- 82% Pearson correlation coefficient, continuous WEB vs 15 item mISA-P r = 0.67 Kappa, dichotomised WEB vs 15 item mISA-P: 0.6	Not stated	Decreased – uses a Likert scoring method for all 10 questions	Measures impact (disempowerment) not specific behaviours
				86	91	52	98	10	41				
Six studies using validation paradigm methods only													
Brown, Lent, Brett, Sas & Pederson (1996)	7 question WAST 2 question WAST-Short	Abuse Risk Inventory (ARI)		In known group comparison, significant difference between abused and non-abused women on total WAST score: 18 vs 9 respectively, p<0.001						Total 7 question WAST score vs total ARI, r=0.96 Individual WAST questions vs ARI, Spearman correlation coefficients, r = 0.80 to 0.85	Cronbach's alpha 0.95 Corrected item total correlations, r = 0.81 to 0.89	Decreased – uses a likert type scale	Decreased - does not explicitly consider sexual IPV
				In known group comparison: WAST-Short ≥ 1 vs ARI***									
				92	100								

Brown, Lent, Schmidt & Sas (2000)	8 question WAST	Abuse Risk Inventory (ARI)								Pearson correlation coefficient, WAST vs ARI: r= 0.69, p=0.01	Cronbach's alpha 0.75	Decreased – uses term abuse and likert type scale	Includes questions on physical, sexual & emotional IPV with impact questions
Brown, Schmidt, Lent, Sas & Lemelin (2001)	8 question French WAST 2 question French WAST-Short	Abuse Risk Inventory (ARI)	Not stated	In known group comparison: French WAST-Short ≥ 1 vs ARI						Total French WAST score vs Total ARI score: r=0.96 Individual French WAST questions vs ARI, r= 0.75 to 0.93	Cronbach's alpha for 8 question French WAST 0.95	Decreased - both use likert type scale	
				79	100								
Chen, Rovi, Washington, Jacobs, Vega, Pan & Johnson (2007)	2 question WAST-Short 4 question HITS	8 question Woman Abuse Screening Tool (WAST)								Total English WAST-Short score vs total 8 question WAST score, 0.81, p<0.001 Total HITS score vs total 8 question WAST score, 0.77, p<0.001	Cronbach's alpha for: WAST-Short 0.8 HITS 0.79	Decreased -both use likert type scale	Decreased – neither explicitly consider sexual IPV
Sagrestano, Rodriguez, Carroll, Bieniarz, Greenberg, Castro & Nuwayhid (2002)	2 questions in Perinatal Self-Administered Inventory (PSAI) English & Spanish versions	CTS subscale on verbal aggression (7items) and physical violence (9 items) English & Spanish versions	In past year: according to CTS verbal aggression 84%, physical violence 17%							First PSAI question vs verbal aggression subscale of CTS, r=0.10 Second PSAI question vs verbal aggression subscale of CTS, r=0.03 Second PSAI question vs physical subscale of CTS, r=- 0.05	Not stated	Decreased – due to complexity of 2 questions (see text)	Decreased - do not explicitly consider sexual IPV Spanish versions of index questions, back translation used
McFarlane, Parker, Soeken, & Bullock (1992)	3 question AAS English & Spanish versions	30 question ISA English & Spanish versions	-							Those positively identified for IPV on the 3 question AAS also scored significantly higher on the ISA (no figures presented)	Not stated		Decreased – no question on emotional IPV. No details on validation of Spanish versions

One study using method from neither paradigm													
Connelly, Newton, Landsverk & Aarons (2000)	Single question in hospital admission protocol	CTS subscale on physical violence (9 items)	9 item CTS: 19% Single question: 4%							Not stated	Not stated		-

*: An optimal data analysis computer program established that a cut off score of 10.5 on the HITS reliably differentiated respondents into two groups.

**: Focus group decided HITS should focus on physical & verbal IPV – though titled as a "Domestic violence screening tool."

***: WAST-Short ≥1: a score of one was assigned to the most extreme positive responses (e.g. "a lot of tension") and a score of zero to other response options.

Table 2: Summary characteristics of 20 studies in systematic review

Study details	Study design	Number of participants	Participants
Peralta & Fleming (2003)	**Index Question:** 1 question: "Do you feel safe at home?	**Eligible:** Not stated	**Age (mean, SD, range):** Mean and SD not stated; range 18-36
	Comparator Tool: 6 questions Modified Conflict Tactics Scale (6 item mCTS) - for physical & / or psychological IPV	**Declining:** 12%	**Ethnicity:** White 61%, Black 26%, Other 13%
	Inclusion criteria: All women within the waiting room of the urban family practice clinic in Madison, Wisconsin who were English-speaking between the age of 18-36 years		
	Exclusion criteria: Non-English speakers	**Recruited:** 399	**Socio-economic status indicators:** More than high school education: 81% Marital status: 41% abused women married, 58% of non-abused women married
	Sample: Abused and non-abused women		
	Type of study: Validation study		
	Setting: USA, urban family practice clinic		
Paranjape, Rask, & Liebschutz (2006)	**Index Questions:** 3 questions: Slapped, Threatened or Thrown	**Eligible:** 324	**Age (mean, SD, range):** 38, 10, not stated
	Comparator Tool: 30 questions Index of Spouse Abuse (ISA)	**Declining:** 84	**Ethnicity:** African American 91%
	Inclusion criteria: Women between 18 - 65 years of age, English speaking, and seen a medical provider within the centre on that day		
	Exclusion criteria: Patients who could not be interviewed alone	**Recruited:** 240	**Socio-economic status indicators:** Median monthly income- $800 Educational level, marital status, housing status, employment and insurance status was collected according to whether participants where positive or negative for IPV. No significant difference seen between 2 groups
	Sample: Abused and non-abused women		
	Type of study: Validation study		
	Setting: USA, urban family practice clinic in an inner city hospital which provides primary care to an impoverished and mostly uninsured population		
Feldhaus, Kozioi-McLain, Amsbury, Norton, Lowenstein & Abbott (1997)	**Index Questions:** 3 questions: Partner Violence screen (PVS)	**Eligible:** 426	**Age (mean, SD, range):** 36, 16, not stated
	Comparator Tool: 30 questions Index of Spouse Abuse (ISA) 16 questions Modified Conflict Tactics Scale (16 item mCTS)	**Declining:** 47, 57 missed due to heavy volume of patients	**Ethnicity:** White 45%, Hispanic 30%, Black 19%, Other 6%
	Inclusion criteria: Non-critical, English speaking women presenting to one of 2 urban ED departments		
	Exclusion criteria: Under the age of 18, had an altered mental status or primary psychiatric disorder	**Recruited:** 322	**Socio-economic status indicators:** Household income <$15,000- 64% Educational level and insurance status
	Sample: Abused and non-abused women		
	Type of study: Validation study		

Study details	Study design	Number of participants	Participants
	Setting: USA, two urban, hospital-based A&E departments		
MacMillan, Wathen, Jamieson, Boyle, McNutt, Woster, Lent & Webb (2006)	**Index Questions:** 3 questions: Partner Violence Screen (PVS) 8 questions: Woman Abuse Screening Tool (WAST)	**Eligible:** 2602	**Age (mean, SD, range):** 37, 12, not stated
	Comparator Tool: 30 question Composite Abuse Scale (CAS)	**Declining:** 141	**Ethnicity:** Born in Canada- 87%
	Inclusion criteria: All women presenting for an appointment at the included sites (EDs, family practices or women's health clinics), aged 18-64 years, at a site for their own health care visit, able to separate themselves from individuals who accompanied them, able to speak and read English, were not too ill to participate and could provide informed consent.		
	Exclusion criteria: Not stated	**Recruited:** 2461	**Socio-economic status indicators:** <$24,000- 18% Woman was the main source of income wages or salary- 58% Educational level, marital status and children living at home
	Sample: Abused and non-abused women		
	Type of study: Validation study- primary aim was to test presentation effects of tools		
	Setting: Canada, Ontario. 2x ED, 2x Family practices, 2x Women's health clinics		
Sohal, Eldridge & Feder (2007)	**Index Questions:** 4 questions: Humiliation, Afraid, Rape, Kick (HARK)	**Eligible:** 429	**Age (mean, SD, range):** 35, 11.95, 18 - 70
	Comparator Tool: 30 question Composite Abuse Scale (CAS)	**Declining:** 197 (14 missed due to heavy volume of patients)	**Ethnicity:** White British 40%, Black British, African or Caribbean 25%, Indian, Pakistani or Bangladeshi 18%
	Inclusion criteria: Women aged more than 17 years waiting to see a doctor or nurse, who had been in an intimate relationship in the last year.		
	Exclusion criteria: Women who were accompanied by children over four years of age or another adult, too unwell to complete the questionnaires, unable to understand English or unable to give informed consent.	**Recruited:** 232	**Socio-economic status indicators:** 51% in paid job, 53% owned house or flat
	Sample: Abused and non-abused women		
	Type of study: Validation study		
	Setting: UK, 12 general practices in a multi-ethnic inner city area of London		
Tiwari, Fong, Chan, Leung, Parker & Ho (2007)	**Index Questions:** 3 individual questions derived from the Chinese Abuse Assessment Screen (AAS)	**Eligible:** 257	**Age (mean, SD, range):** 36, 8, not stated
	Comparator Tool: 3 subscales of physical, emotional & sexual IPV of the 39 question Chinese revised Conflict Tactics Screen (CTS2)	**Declining:** 0	**Ethnicity:** All Chinese

Study details	Study design	Number of participants	Participants
	Inclusion criteria: Not stated		
	Exclusion criteria: Not stated	**Recruited:** 100 pregnant women and 157 non-pregnant women.	**Socio-economic status indicators:** 91% married women 44% monthly family incomes lower than the official median of HK$11,000 (about US$1,375) 35% less than 10 yrs of schooling
	Sample: Abused and non-abused pregnant and non-pregnant women		
	Type of study: Validation study		
	Setting: Hong Kong, antenatal clinic of a public hospital and a community centre		
Reichenheim & Moraes (2003)	**Index Questions:** 1 question Portuguese Abuse Assessment Screen's (AAS) anchor question on physical IPV during pregnancy	**Eligible:** 3800	**Age (mean, SD, range):** 24, 6, range not stated
	Comparator Tool: 12 question physical violence scale of modified revised Portuguese Conflict Tactics Screen (12 item mCTS2)	**Declining:** 3	**Ethnicity:** All Brazilian, Portuguese speaking
	Inclusion criteria: Given birth within 24 hours; interviews conducted in first 48 hours postpartum. Included all premature births within six month period		
	Exclusion criteria: Diabetes mellitus, systematic arterial hypertension, given birth to neonates with severe congenital malformations, infections associated with prematurity, or twins. Not in steady relationships involving current or former partners	**Recruited:** 748	**Socio-economic status indicators:** Median monthly income per capita of US$97 (95%CI 26 to 346) 57% attended less than 8 yrs of school 75% either married or living with a partner at the time 6 prenatal visits on average
	Sample: Abused and non-abused post-natal women		
	Type of study: Validation study		
	Setting: Brazil, Rio de Janeiro. Three public sector maternity wards		
Ernst, Weiss, Cham, Hall & Nick (2004)	**Index Questions:** 4 question Ongoing Violence Assessment Tool (OVAT)	**Eligible:** 362	**Age (mean, SD, range):** 34, 10, range not stated
	Comparator Tool: 30 question Index of Spouse Abuse (ISA)	**Declining:** 46 (10 did not complete forms)	**Ethnicity:** Caucasian 49%, African American 16%, Hispanic 20%, Asian or other 15%
	Inclusion criteria: English speaking patients entering ED department		
	Exclusion criteria: Under the age of 18, had no current partner, had an altered mental state, had an underlying psychiatric diagnosis, were too ill to participate or drug or alcohol intoxicated	**Recruited:** 212 women & 94 men.	**Socio-economic status indicators:** 41% married, 57% not married, 2% not stated 66% with children, 33% without children, 1% not stated
	Sample: Abused and non-abused women		
	Type of study: Validation study		
	Setting: USA, A&E department		

Study details	Study design	Number of participants	Participants
Chen, Rovi, Vega, Jacobs & Johnson (2005)	**Index Questions:** 4 questions - Hurts, Insults, Threatens and Screams (HITS). English & Spanish versions	**Eligible:** 386	**Age (mean, SD, range):** 36, SD & range not stated
	Comparator Tool: 11 question Index of Spouse Abuse-Physical dimension (ISA-P) 8 question Woman Abuse Screening Tool (WAST) English & Spanish versions	**Declining:** 128 refused, 56 did not complete questionnaire due to long waiting period for a private room	**Ethnicity:** Hispanic 72%, non-Hispanic White 20%, non-Hispanic Black 6%, non-Hispanic Other 1%. Country of origin for Hispanics: 39% Cuban / Cuban American, 35% Puerto Rican, 11% Dominican, 5% Mexican / Mexican American, 10% other Latin American 44% completed interviews in Spanish
	Inclusion criteria: Women attending an urban family practice site who were 18 years or older and were currently involved in an ongoing relationship		
	Exclusion criteria: Not stated	**Recruited:** 202	**Socio-economic status indicators:** Total mean income $10,757 For English speaking $14,142 For Spanish speaking $6,461 Significant differences between those who carried out the study in English & those in Spanish, including between Hispanics. No significant differences between Hispanics and non-Hispanics
	Sample: Abused and non-abused women		
	Type of study: Validation study		
	Setting: USA, urban family practice site		
Sherin, Sinacore, Li, Zitter & Shakil (1998)	**Index Questions:** 4 questions - Hurts, Insults, Threatens and Screams (HITS)	**Eligible:** Not stated	**Age (mean, SD, range):** Not stated
	Comparator Tool: 15 question modified Conflict Tactics Scale (15 item mCTS) – verbal & physical aggression items	**Declining:** Not stated	**Ethnicity:** Not stated
	Inclusion criteria: Phase 1- Patients at Family Practice Centre, aged 21 or over and lived with the same partner for at least 12 months. Phase 2 - Self-identified women who had experienced IPV, recruited via a crisis shelter or an emergency room		
	Exclusion criteria: Not stated	**Recruited:** Phase 1: 160 women recruited from general practice; Phase 2: 99 self-identified survivors of partner violence, (54 via shelter; 45 via emergency room)	**Socio-economic status indicators:** Not stated
	Sample: Abused and non-abused women; in phase 2 a known group comparison		
	Type of study: Validation study & known group comparison		
	Setting: USA. Phase 1 - Family practice. Phase 2 - IPV crisis shelter and an emergency room		

Study details	Study design	Number of participants	Participants
Zink, Levin, Putnam & Beckstrom (2007)	**Index Questions:** 5 non-graphic questions that can be used when children are present	**Eligible:** 450	**Age (mean, SD, range):** Not stated
	Comparator Tool: 39 question revised Conflict Tactic Scale (CTS2)	**Declining:** 50 refused participation, 7 data not analysed (5 answered every question with 0)	**Ethnicity:** African American 51%, White 49%
	Inclusion criteria: English speaking mothers, in a relationship with a steady partner for at least 1 year and with at least 1 child between 3 & 12 yrs of age.		
	Exclusion criteria: Not stated	**Recruited:** 393	**Socio-economic status indicators:** Income $40,000/yr 31%, $20,000-$40,000/yr 34%, <$20,000 34% >12th grade 40%, ≤ 12th grade 60% Married 81%, Single 13%, Separated / divorced 6%
	Sample: Abused and non-abused women		
	Type of study: Validation study		
	Setting: USA, Cincinnati, Ohio. 5 paediatric and family medicine practices.		
Bonomi, Thompson, Anderson, Rivara, Holt, Carrell & Martin (2006)	**Index Questions:** 5 behavioural tactic abuse questions - Behavioural Risk factor Surveillance Survey (BRFSS)	**Eligible:** 2,504	**Age (mean, SD, range):** 46, 12, range not stated
	Comparator Tool: 10 impact questions - Women's Experience with Battering scale (WEB)	**Declining:** 0	**Ethnicity:** White 83%, Hispanic 4%, No information on 13%
	Inclusion criteria: Women enrolled for at least 3 years in a Group Health Cooperative, aged between 18-64 years		
	Exclusion criteria: Women who have never had an intimate partner or who resided outside of Washington State	**Recruited:** 2,504	**Socio-economic status indicators:** >$75,000 - 34%, $50,000- $75,000- 27%, $25,000- $50,000- 28%, <$25,000- 11% 81% employed, 87% completed some college or more 65% married, 34% children in home
	Sample: Abused and non-abused women		
	Type of study: Validation study		
	Setting: USA, Washington State. Telephone survey of randomly selected women		
Coker, Pope, Smith, Sanderson & Hussey (2001)	**Index Questions:** 10 impact questions - Women's Experience with Battering scale (WEB)	**Eligible:** 1503	**Age (mean, SD, range):** 38, 11, range not stated
	Comparator Tool: 15 question modified Index of Spouse Abuse- Physical (15 item mISA-P)	**Declining:** 174 refused, 97 did not complete health assessment interview, 80 had missing data on several response variables	**Ethnicity:** African American 62%, White 38%
	Inclusion criteria: Women seeking medical care in a family practice clinic, aged between 18 and 65, were insured by a managed care organisation and/or Medicaid and ever been in an intimate, sexual relationship with a man for a least 3 months		
	Exclusion criteria: Women whose partners would not leave them alone were not recruited	**Recruited:** 1152	**Socio-economic status indicators:** 78% insured by a managed care provider, 22%

Study details	Study design	Number of participants	Participants
	Sample: Abused and non-abused women		insured by Medicaid 33% college graduates, 56% high school graduate or some college, 11% less than high school 86% were currently employed 39% married, 35% single, 21% divorced / separated, 5% widowed
	Type of study: Validation study		
	Setting: USA, two university associated family practice clinics		
Brown, Lent, Brett, Sas & Pederson (1996)	**Index Questions:** 7 questions Woman Abuse Screening Tool (7 item WAST) 2 question WAST-Short (= 2 questions from the 7 item WAST that women were most comfortable with)	**Eligible:** Out of comparison group: 38 women approached In abuse group: unknown	**Age (mean, SD, range):** Abused: 32, SD not stated, 18 – 57, Non-abused: 42, SD not stated, 25 - 61
	Comparator Tool: Abuse Risk Inventory (ARI)	**Declining:** Out of comparison group 11 refused, 2 did not appear for interview and 1 during interview was "identified as having experienced abuse in her current relationship; her data was not included in the analysis." In abuse group: unknown	**Ethnicity:** Not stated
	Inclusion criteria: Women staying at local shelter due to abuse by male partners		
	Exclusion criteria: All comparison groups subjects were asked not to participate if they had a history of spousal abuse	**Recruited:** 48 (24 abused, 24 non-abused)	**Socio-economic status indicators:** Abused / Non-abused Employed: 25% / 100% > Can$30,000: 33% / 100% Married: 12% / 79% Separated / Divorced 50% / 4% Single 38% / 17% College / University Education: 22% / 74%
	Sample: Known group analysis using intentionally selected women representing 2 extreme groups of abused women at a local shelter and non-abused women, recruited from the principal investigator's professional contacts		
	Type of study: Validation study, using a known group comparison		
	Setting: Canada, Western Ontario. At women's shelter		
Brown, Lent, Schmidt & Sas (2000)	**Index Questions:** 8 questions Woman Abuse Screening Tool (WAST)	**Eligible:** 399 patients; 44 physicians	**Age (mean, SD, range):** 46; SD not stated; 18-86
	Comparator Tool: Abuse Risk Inventory (ARI)	**Declining:** 92 patients; 24 physicians	**Ethnicity:** White 98%

Study details	Study design	Number of participants	Participants
	Inclusion criteria: Using a stratified random sampling frame, 20 physicians needed to be selected from 400 in London, Ontario, Canada -) Women needed to be 18 or older, attending for a periodic health examination, for prenatal care or acute symptoms of illness, be English speaking, unaccompanied by another person, currently involved in an intimate relationship (married or common law) and they had to consider the attending physician their primary care physician.		
	Exclusion criteria: Not stated **Sample:** Abused and non-abused women **Type of study:** Validation study **Setting**: Canada, South Western Ontario, Family practices – urban and rural	**Recruited:** 307 patients; 20 physicians	**Socio-economic status indicators:** Employed 59% Income >$30,000 - 59% Married or in common law relationship 88%
Brown, Schmidt, Lent, Sas & Lemelin (2001)	**Index Questions:** 8 questions French Woman Abuse Screening Tool (WAST) 2 question French WAST-Short	**Eligible:** Not stated	**Age (mean, SD, range):** SD not stated Mean:- Abused 38 Non-abused 36 Range:- 27-54 17-58
	Comparator Tool: Abuse Risk Inventory (ARI) **Inclusion criteria:** 18 years of age (not adhered to) In a couple relationship for the last 12 months	**Declining:** Not stated	**Ethnicity:** French speaking women
	Exclusion criteria: Not stated **Sample**: Abused and non-abused women **Type of study:** Validation study, using a known group comparison **Setting:** Canada, Ontario & Quebec - refuge and private homes	**Recruited:** 46 25 abused 21 non-abused	**Socio-economic status indicators:** Abused Non-abused Employed: 9% 92% Can$30,000: 15% 95% Married: 32% 81%
Chen, Rovi, Washington, Jacobs, Vega & Pan (2007)	**Index Questions:** 2 question English WAST-Short 4 question - Hurts, Insults, Threatens and Screams (HITS). English version	**Eligible:** 730	**Age (mean, SD, range):** 36, SD & range not stated
	Comparator Tool: 8 question English Woman Abuse	**Declining:** 200 refused to	**Ethnicity:** African-American 71%, Hispanic

Study details	Study design	Number of participants	Participants
	Screening Tool (WAST)	participate, 7 did not complete the questionnaire because of the waiting time for a private room	14%, White 12%, Other 4%
	Inclusion criteria: aged 18 years or older, currently involved with a partner		
	Exclusion criteria: “”	**Recruited:** 523	**Socio-economic status indicators:** 29% completed college Mean income $20,423 73% employed, including part time work
	Sample: Abused and non-abused women		
	Type of study: Validation study		
	Setting: – 4 urban family medicine practices		
Sagrestano, Rodriguez, Carroll, Bieniarz, Greenberg, Castro & Nuwayhid (2002)	**Index Questions:** 2 questions on IPV within the Perinatal Self-Administered Inventory (PSAI)	**Eligible:** 196	**Age (mean, SD, range):** 25.7, 6.0, 14-41
	Comparator Tool: Seven questions on verbal aggression & nine on physical violence from the Conflict Tactics Scale (CTS)	**Declining:** 0	**Ethnicity:** African American 48%, Hispanic 46%, White or other 6% 25% (n=42) completed survey in Spanish
	Inclusion criteria: Women in a waiting room of a women’s care centre scheduled for routine, antenatal care		
	Exclusion criteria: Less than 20 weeks pregnant; were accompanied by small children who could not leave the waiting room with another relative; did not speak English or Spanish	**Recruited:** 196, but only 166 entered into analysis	**Socio-economic status indicators:** Median annual income was $10,000 to $20,000 49.4% earned less than$10,000
	Sample: Abused and non-abused women		
	Type of study: Validation study		
	Setting: Mid-Western USA, university affiliated women’s care centre		
McFarlane, Parker, Soeken, & Bullock (1992)	**Index Questions:** 3 question Abuse Assessment Screen AAS	**Eligible:** 691	**Age (mean, SD, range):** Age ranged from 13 to 30+ years (13 to 19 years- 31%, 20 to 29 years- 57%, >30 years- 12%); -, -
	Comparator Tool: 30 question Index of Spouse Abuse (ISA)	**Declining:** 0	**Ethnicity:** Black- 39%, Hispanic- 34% (most Mexican American), White- 27% Survey completed in English and Spanish – numbers in each language group not known
	Inclusion criteria: Attending one of two prenatal clinics		
	Exclusion criteria: Not stated	**Recruited:** 691	**Socio-economic status indicators:** 95% below poverty level (not defined)
	Sample: Abused and non-abused women		

Study details	Study design	Number of participants	Participants
	Type of study: Validation study		35% married
	Setting: US, Texas, Houston & Baltimore; two prenatal clinics		
Connelly, Newton, Landsverk & Aarons (2000)	**Index Questions:** 1 question: "Are you in a relationship in which you have been threatened, scared or hurt by someone?" If yes, whom? (Part of hospital admission protocol)	**Eligible:** Not stated	**Age (mean, SD, range):** 23, 6.2, 14-42 46% were 21 years or younger
	Comparator Tool: 9 question physical subscale of the Conflict Tactics Scale (CTS)	**Declining:** Not stated	**Ethnicity:** Hispanic 40% Caucasian 27% African American 23% Asian, Pacific Islander, Native American/other- 9%
	Inclusion criteria: Mothers giving birth between Feb 1996 and Mar 1997 who participated in a randomised clinical trial of paraprofessional home visitation services. High risk mothers were identified 24 hours after birth using 15-item screen. Participants had to be English or Spanish Speaking, not active to child protective services and referenced "baby's father" for the CTS.		
	Exclusion criteria: Not stated	**Recruited:** 436	**Socio-economic status indicators:** 53% not completed high school 21% only completed high school 84% not married 53% reported father of the baby lived in the home
	Sample: Abused and non-abused high risk post partum mothers		
	Type of study: Validation study, part of larger trial		
	Setting: San Diego, US. Unclear whether all questions administered at hospital or home		

Table 3: Correlation measures and their interpretation (for criterion correlation validity, convergent validity and association between index scores & external variables)

STUDIES	REFERENCE STANDARD	CORRELATION COEFFICIENT	ACTUAL VALUE	PAPER'S INTERPRET -ATION
Tiwari et al, 2007 3 individual questions derived from the Chinese AAS Qs	Yes	Kappa coefficient	Physical IPV 0.56 Emotional IPV 0.52 Sexual IPV 0.47	Fair agreement
Reicheheim et al, 2004 Portuguese anchor AAS Q.	Yes	Point biserial correlation	0.68	High
Ernst et al, 2004 OVAT	Yes	Kappa statistic	0.58	-
Chen et al, 2005 English HITS vs English ISA-P English HITS vs English WAST Spanish HITS vs Spanish ISA-P Spanish HITS vs Spanish WAST	Not stated	-	 0.76 0.76 0.81 0.78	-
Sherin et al, 1998 HITS	Not stated	-	0.85	-
Bonomi et al, 2006 BRFSS	No	Numbers & percentages of women overlapping between BRFSS and WEB	BRFSS+/WEB+ 126=5% BRFSS+/WEB- 240=9% BRFSS-/WEB+ 48=2% BRFSS-/WEB- 2085=83%	-
Coker et al, 2001 WEB	No	Numbers & percentages of women overlapping between WEB & ISA-P Pearson correlation coefficient Cohen's kappa statistic Association with self reported poor mental health	WEB+/ISA-P+ 98=9% WEB+/ISA-P- 92=8% WEB-/ISA-P+ 16=1% WEB-/ISA-P- 946=82% 0.67 60% Relative risk 6.25, 95% CI 2.72 – 14.32	- Good agreement Good agreement Strong association

STUDIES	REFERENCE STANDARD	CORRELATION COEFFICIENT	ACTUAL VALUE	PAPER'S INTERPRET -ATION
Coker et al, 2001 WEB Continued		Association with ≥ 10 physician visits in last year	Relative risk 1.05, 95% CI 0.82 – 1.37	-
Brown et al, 1996 Total 7 Q.WAST score Individual WAST Qs	No	 - Spearman correlation coefficient	 0.96 0.80 to 0.85	
Brown et al, 2000 8 Q. WAST	No	Pearson correlation coefficient	0.69	
Brown et al, 2001 Total French 8 Q. WAST Individual French WAST Qs	No	 - -	 0.96 0.75 to 0.93	
Chen et al, 2007 English WAST-Short HITS	No		 0.81 0.77	 Highly correlated Highly correlated
Sagrestano et al, 2002 First PSAI Q vs CTS (verbal agression subscale) Second PSAI Q vs CTS (verbal agression subscale) Second PSAI Q vs CTS (physical agression subscale)	No	**-**	 0.1 0.03 0.5	 No Correlation
McFarlane et al, 1992 3 AAS Qs	No	-	No data presented	Valid and specific in identifying abuse

Table 4: Internal consistency reliability measures and their interpretation

STUDIES	CRONBACH'S ALPHA	CORRECTED ITEM-TOTAL CORRELATIONS	PAPER'S INTERPRETATION	UNIDIMENSIONAL CONSTRUCT
Ernst et al, 2004 OVAT	0.6		Passable	No
Chen et al, 2005 English HITS Spanish HITS	0.76 0.61			No No
Sherin et al, 1998 English HITS	0.8			No
Zink et al, 2007 Five non-graphic DV questions	0.46		Mediocre	No
Brown et al, 1996 Seven question WAST	0.95	0.81 – 0.89	High	No
Brown et al, 2000 Eight question WAST	0.75			No
Brown et al, 2001 French WAST-Short	0.95		Good	Yes
Chen et al, 2007 English WAST-Short English HITS	0.80 0.79		Good Good	Yes No

Table 5: QUADAS quality items

STUDIES	**No.1[1] Spectrum of patients represent-ative**	**No.2[2] Inclusion criteria stated**	**No.3[3] Accept-able reference standard**	**No.4[4] Time period between tools short enough**	**No.5[5] Sample verified with refer-ence standard**	**No.6[6] All receive same refer-ence standard**	**No.7[7] Reference standard independ-ent of index tool**	**No.8[8] Enough detail to replicate index tool**	**No.9[9] Enough detail to replicate reference standard**	**No.10[10] Blind analysis of index tool**	**No.11[11] Blind analysis of refer-ence standard**	**No.12[12] Same clinical data available in practice**	**No.13[13] Un-interpret-able / intermed-iate results presented**	**No.14[14] With-drawals from study explained**	**No. of items fulfilled**
Five diagnostic accuracy studies															
Peralta et al, 2003 Safety Question	Yes	Yes	**No**	Yes	Yes	Yes	Yes	Yes	Yes	Unknown	Unknown	Yes	No	Yes	10
Paranjape et al, 2006 STaT	Yes	Yes	Yes	Yes	Yes	Yes	**No**	Yes	Yes	Unknown	Unknown	Yes	Yes	Yes	11
Feldhaus et al, 1997 PVS	Yes	Yes	ISA: Yes mCTS **No**	Yes	Yes	Yes	Yes	Yes	Yes	Unknown	Unknown	Yes	Yes	Yes	11
MacMillan et al, 2006 PVS & WAST	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Unknown	Unknown	Yes	**No**	Yes	11
Sohal et al, 2007 HARK	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	14

STUDIES	**No.1 Spectrum of patients represent-ative**	**No.2 Inclusion criteria stated**	**No.3 Accept-able reference standard**	**No.4 Time period between tools short enough**	**No.5 Sample verified with referen-ce stan-dard**	**No.6 All receive same referen-ce stan-dard**	**No.7 Reference standard independ-ent of index tool**	**No.8 Enough detail to replicate index tool**	**No.9 Enough detail to replicate reference standard**	**No.10 Blind analysis of index tool**	**No.11 Blind analysis of reference standard**	**No.12 Same clinical data available in practice**	**No.13 Un-interpret able / intermedi ate results presented**	**No.14 Withdra-wals from study explained**	**No. of items fulfilled**
Eight studies using diagnostic accuracy and validation paradigms															
Tiwari et al, 2007 Chinese 3 AAS Qs	Yes	**No**	Yes	Yes	Yes	Yes	Unclear	Yes	Unclear	Unknown	Unknown	Yes	Yes	Yes	9
Reicheheim et al, 2004 Portuguese 1 AAS Q.	Yes	-	Yes	Yes	Yes	Yes	Unclear	No	Yes	Unknown	Unknown	Yes	Yes	Yes	9
Ernst et al, 2004 OVAT	Yes	Yes	Yes	Yes	Yes	Yes	**No**	Yes	Yes	Unknown	Unknown	Yes	Yes	**No**	10
Chen et al, 2005 English & Spanish HITS	Yes	Yes	Unclear X4 reference standards	Yes	Yes	**No**	**No**	Yes	Yes	Unknown	Unknown	Yes	Yes	Yes	9
Sherin et al, 1998 HITS	**No**	Yes	**No**	Yes	Yes	Yes	Yes	Yes	Yes	Unknown	Unknown	Yes	Yes	**No**	9
Zink et al, 2007 5 DV Qs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Unknown	Unknown	Yes	Yes	Yes	12
Bonomi et al, 2006 BRFSS	Yes	Yes	**No**	Yes	Yes	Yes	Yes	Yes	Yes	Unknown	Unknown	No	Yes	Yes	10
Coker et al, 2001 WEB	Yes	Yes	Unclear	Yes	Yes	Yes	Yes	Yes	Yes	Unknown	Unknown	Yes	Yes	Yes	11

STUDIES	**No.1 Spectrum of patients represent-ative**	**No.2 Inclusion criteria stated**	**No.3 Accept-able reference standard**	**No.4 Time period between tools short enough**	**No.5 Sample verified with referen-ce stan-dard**	**No.6 All receive same referen-ce stan-dard**	**No.7 Reference standard independ-ent of index tool**	**No.8 Enough detail to replicate index tool**	**No.9 Enough detail to replicate reference standard**	**No.10 Blind analysis of index tool**	**No.11 Blind analysis of reference standard**	**No.12 Same clinical data available in practice**	**No.13 Un-interpret able / intermedi ate results presented**	**No.14 Withdra-wals from study explained**	**No. of items fulfilled**
Six studies using validation paradigm methods only															
Brown et al, 1996 7 Q. WAST	**No**	Yes	Unclear	Yes	Yes	Yes	Yes	Yes	No	Unknown	Unknown	Yes	Yes	Yes	9
Brown et al, 2000 8 Q. WAST	Yes	Yes	Unclear	Yes	Yes	Yes	Yes	Yes	No	Unknown	Unknown	Yes	Yes	No	9
Brown et al, 2001 French 8 Q. WAST & WAST-S	**No**	-	Unclear	-	-	-	Yes	Yes	No	Unknown	Unknown	-	-	-	-
Chen et al, 2007 English WAST-Short, HITS	Yes	Yes	**No**	Yes	Yes	Yes	**No**	Yes	Yes	Unknown	Unknown	Yes	Yes	Yes	10
Sagrestano et al, 2002 2 PSAI Qs	Yes	Yes	**No**	No	Yes	Yes	No	Yes	Yes	Unknown	Unknown	Yes	Yes	Yes	9
McFarlane et al, 1992 3 AAS Qs	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Unknown	Unknown	Yes	No	Yes	10
One study using neither research paradigm															
Connelly et al, 2000 Single Q	No	Yes	No	**No**	Yes	Yes	Yes	Yes	Yes	Unknown	Unknown	Yes	No	Yes	8

*Important results are highlighted in bold

[1] Item No. 1) Spectrum of patients' representative?

[2] Item No. 2) Inclusion criteria stated?

[3] Item No. 3) Acceptable reference standard?
[4] Item No. 4) Time period short enough between administered tools?
[5] Item No. 5) Whole / random selection of sample verified with reference standard?
[6] Item No. 6) All participants receive the same reference standard?
[7] Item No. 7) Reference standard independent of index tool? (did not form part of reference standard)
[8] Item No. 8) Enough detail to replicate execution of index tool?
[9] Item No. 9) Enough detail to replicate execution of reference standard?
[10] Item No. 10) Blind analysis of index tool?
[11] Item No. 11) Blind analysis of reference standard?
[12] Item No. 12) Same clinical data available when interpreted as would be available in practice?
[13] Item No. 13) Un-interpretable / intermediate results presented?
[14] Item No. 14) Withdrawals from the study explained?

Table 6: Ethnicity quality criteria

STUDIES	IS ETHNICITY DESCRIBED?	TERMS USED TO DESCRIBE ETHNICITY:	IS CLASSIFICATION SYSTEM USING ETHNICITY JUSTIFIED?	IS ETHNICITY SELF-ASSIGNED?	ARE SOCIO-ECONOMIC FACTORS CONSIDERED?
Peralta et al, 2003 Safety Question	Yes	"Racial identity" "Ethnic differences" – used in Discussion	No	Yes	Yes
Paranjape et al, 2006 STaT	Yes	-	No	Unclear	-
Feldhaus et al, 1997 PVS	Yes	"Racial or ethnic"	No	Yes	-
MacMillan et al, 2006 PVS & WAST	Yes	"Born in Canada"	No	Unclear	-
Sohal et al, 2007 HARK	Yes	"Ethnic origin" - based on national census categories	Representativeness checked	Yes	-
Tiwari et al, 2007 Chinese 3 AAS Qs	Uncertain	"Chinese" (?ethnicity / language / nationality)	No	Unclear	-

STUDIES	IS ETHNICITY DESCRIBED?	TERMS USED TO DESCRIBE ETHNICITY	IS CLASSIFICATION SYSTEM USING ETHNICITY JUSTIFIED?	IS ETHNICITY SELF-ASSIGNED?	ARE SOCIO-ECONOMIC FACTORS CONSIDERED?
Reicheheim et al, 2004 Portuguese 1 AAS Q.	Yes	"Portuguese speaking in Brazil" (Language)	No	-	-
Ernst et al, 2004 OVAT	Yes	"Race" – including "Caucasian."	No	Yes	-
Chen et al, 2005 English & Spanish HITS	Yes	"Race / ethnicity." (Language spoken and country of origin was described for the Hispanic population).	No	Unclear	Yes
Sherin et al, 1998 HITS	No	-	-	-	-
Zink et al, 2007 5 DV Qs	Yes	"Ethnicity/race" & "Ethnicity"	No	Unclear	Yes
Bonomi et al, 2006 BRFSS	Yes	"Race / ethnicity."	No	Yes	-
Coker et al, 2001 WEB	Yes	"Race"	No	Unclear	-

STUDIES	IS ETHNICITY DESCRIBED?	TERMS USED TO DESCRIBE ETHNICITY:	IS CLASSIFICATION SYSTEM USING ETHNICITY JUSTIFIED?	IS ETHNICITY SELF-ASSIGNED?	ARE SOCIO-ECONOMIC FACTORS CONSIDERED?
Brown et al, 1996 7 Q. WAST	No	-	-	-	-
Brown et al, 2000 8 Q. WAST	Yes	-	No	Unclear	-
Brown et al, 2001 French 8 Q. WAST & WAST-S	Yes	"Francophone" (Language)	No	-	-
Chen et al, 2007 English WAST-Short & HITS	Yes	"Race / ethnicity"	No	Yes	-
Sagrestano et al, 2002 2 PSAI Qs	Yes	"Multiethnic women" & "ethnic minority women"	No	Unclear	-
McFarlane et al, 1992 3 AAS Qs	Yes	"Ethnic or racial"	No	Yes	No
Connelly et al, 2000 Single Q.	Yes	No specific terms used though categories included "Caucasian," "Hispanic" and "Asian"	No	Unclear	-

I now present my narrative results. I have grouped studies together according to the methods used to obtain evidence of validation. The first group are the five studies that reported exclusively diagnostic accuracy. The second group are the eight studies that reported validation paradigm methods and which also made use of diagnostic accuracy. Immediately after these first two groups' narrative results, I have presented an analysis of their collective results. This overview allows a clear interpretation of what the study results actually mean. The third group of studies are the six studies that exclusively reported validation paradigm methods. Following this third group's narrative results, I have examined the meaning of the correlation measures which appear consistently in the studies in this third group and a number of the other systematic review studies. This includes an analysis of the effect of the heterogeneity of the study population on these correlation measures.

3.2.1. Five studies reporting exclusively diagnostic accuracy (i.e. criterion related concurrent validity)

3.2.1.1. Single Safety Question

Peralta and Fleming,[208] compared the single question: "Do you feel safe at home?" to identify IPV to a modified six question version of the Conflict Tactic Scale (6 item mCTS) in 399 English speaking women attending urban family practice in the US. They were 61% white, 26% African-American and 13% other. The validation of the safety question was not compared between these groups.

Quality Appraisal

The 6 item mCTS was not an acceptable reference standard. It only contained six items from the original CTS instead of 19. Five of these items related to psychological violence and only one to physical violence. It had not undergone a validation process (unlike the original CTS or the CTS 2). The terms "African American" and "black" were used interchangeably as well as the terms "racial identity" and "ethnicity." The

"other" grouping included women of Hispanic, Asian and Native American descent. It was justified on the basis that there were too few numbers participating for meaningful analysis to be conducted.

3.2.1.2. Slapped, Threatened or Thrown (STaT)

Paranjape and colleagues,[209] evaluated the three question STaT against the 30 question Index of Spouse Abuse (ISA), in 240 women attending an urgent care centre in south-eastern US who were 91% African-American and all English speaking. This study was in effect looking at the validation evidence for the STaT questions in one ethnic group. There was no assessment of STaT's validity in any other ethnic group apart from the African-Americans. Therefore no comparison can be made of whether STaT's validity varies between ethnic groups.

Box 2: STaT (Slapped, Threatened or Thrown) questions

1. Have you ever been in a relationship where your partner has pushed or slapped you?
2. Have you ever been in a relationship where your partner threatened you with violence?
3. Have you ever been in a relationship where your partner has thrown, broken or punched things?

Quality Appraisal

The reference standard (ISA) was not completely independent of the index tool (STaT) as some of the questions overlapped. Though the ISA has been generally recognised as an acceptable reference standard, [226] I would suggest that as it contains only one question about sexual abuse it is not a perfect gold standard for IPV. Socio-economic data was collected in this African-American sample with a

conclusion that STaT's diagnostic properties could only be generalised to similar patient populations.

3.2.1.3. Partner violence screen (PVS) – two studies

Two studies investigated the PVS which contains one question about physical IPV and two questions about safety.

The three question PVS was tested against two comparators, the thirty question ISA and the modified 16 question version of the Conflict Tactic Scale (16 item mCTS).[210] This comprised the verbal aggression and violence scales of the original CTS. Both comparators were being used as criterion standards by Feldhaus and colleagues in an urban accident and emergency department setting. The 278 women who stayed for the CTS scales were 45% white, 30% Hispanic, 19% black and 6% other. All were English speaking. The PVS's diagnostic accuracy was not tested individually in specific ethnic groups.

Box 3: Partner Violence Screen (PVS) questions

1. Have you been hit, kicked, punched, or otherwise hurt by someone within the past year? If so, by whom?
2. Do you feel safe in your current relationship?
3. Is there a partner from a previous relationship who is making you feel unsafe now?

The single question about physical IPV was more sensitive and specific than the two questions about safety, functioning very similarly to the full PVS.

Quality Appraisal

The 16 item mCTS is not an acceptable reference standard as in its modified state its validity for identifying IPV has not been assessed by any preceding empirical research. Additionally it does not have a question specific to sexual IPV.

MacMillan and colleagues,[211] reported a validation of the three question PVS and the eight question Woman Abuse Screening Tool (WAST), using the thirty question Composite Abuse Scale (CAS), a relatively new comparator, in 2,461 women attending either one of two accident and emergency departments, two family practices or two women's health clinics. The CAS was described as the criterion standard. 87% of the total study population was born in Canada. All were able to speak and read English.

Box 4: Woman Abuse Screening Tool (WAST) questions

1. In general, how would you describe your relationship?

□ A lot of tension

□ Some tension

□ No tension

2. Do you and your partner work out arguments with:

□ Great difficulty?

□ Some difficulty?

□ No difficulty?

3. Do arguments ever result in you feeling down or bad about yourse1f?

□ Often

□ Sometimes

□ Never

4. Do arguments ever result in hitting, kicking or pushing?

□ Often

□ Sometimes

□ Never

5. Do you ever feel frightened by what your partner says or does?

□ Often

□ Sometimes

□ Never

6. Has your partner ever abused you physically?

□ Often

□ Sometimes

□ Never

7. Has your partner ever abused you emotionally?

□ Often

□ Sometimes

□ Never

8. Has your partner ever abused you sexually?

□ Often

□ Sometimes

□ Never

Quality Appraisal:
87% of the total study population was born in Canada. This corresponds closely to their nationality.

3.2.1.4. HARK

Sohal and colleagues,[200] validated the four question HARK which was developed from the Abuse Assessment Screen (AAS). The HARK was compared to the 30 question CAS, (the same comparator that was used in the MacMillan study above) in 232 women attending UK general practice. The study population was ethnically diverse but the HARK questions' validity was not compared between ethnic groups.

Box 5: HARK (Humiliation, Afraid, Rape, Kick) questions

1 **H**UMILIATION
Within the last year, have you been humiliated or emotionally abused in other ways by your partner or your ex-partner?

2 **A**FRAID
Within the last year, have you been afraid of your partner or ex-partner?

3 **R**APE
Within the last year, have you been raped or forced to have any kind of sexual activity by your partner or ex-partner?

4 **K**ICK
Within the last year, have you been kicked, hit, slapped or otherwise physically hurt by your partner or ex-partner?

Quality Appraisal:
This is the only study that had blind analysis of the index tool and the reference standard in that when the researcher totalled a participant's score on the four HARK questions she did not know the individual's CAS score and vice versa. National census categories were used. 40% of the study population described their ethnic origin as white British, 25% as black British, African or Caribbean and 18% as Indian, Pakistani or Bangladeshi.

The ethnicity of the study population was compared to the local population in the London borough of Newham by using the National Census 2001 figures. This comparison neatly showed that the percentage of the study population that described their ethnic origin as white British was 40% (6% higher than that in the local population according to the census, i.e. 34%) whilst 18% of the study population described their ethnic origin as Indian, Pakistani or Bangladeshi (11% lower than that in the local population according to the census i.e. 29%). The authors concluded that this analysis showed that the study population was not representative of the local population. This lack of representativeness would not have necessarily affected the sensitivity or specificity calculations unless the women who did not take part in the study (i.e. the missing south Asian women) answered differently with regards to only one of the instruments (the HARK or the CAS). If these missing south Asian women affected the prevalence of IPV (either increasing or decreasing it) then this may have had an effect on the PPV and NPV for the HARK. Comparison to the local population revealed that the study population was of a higher socio-economic status, as reflected by the higher percentage in a paid job (12% higher) and owning a house or flat (9% higher).[200]

I now consider the results of these five studies collectively. This allows a clear interpretation of what the study results actually mean.

3.2.1.5. Overview of five studies reporting exclusively diagnostic accuracy

Out of these five studies that exclusively reported diagnostic accuracy data,[200, 208-211] one did not use an acceptable reference standard which makes it impossible to interpret the poor diagnostic accuracy indices generated.[208] Out of the four remaining studies, two did not attempt to identify sexual IPV [209, 210] which was not reflected in the diagnostic indices as the reference standard (ISA) only contained one question about sexual IPV. Out of the final two studies,[200, 211] , the questions in one[200] resulted in a far greater change from pre-test to post-test probability (60%) than the questions in the other (37% for PVS and 45% for eight question WAST).[211] Additionally, the HARK questions have a simple scoring system whereas the WAST's scoring is more complex (uses a likert type scale), potentially affecting response processes. The HARK questions also have good content validation with separate questions on physical, sexual and emotional IPV.

I will now consider the eight studies reporting validation paradigm methods which also made use of diagnostic accuracy.

3.2.2. Eight studies reporting validation paradigm methods with diagnostic accuracy

3.2.2.1. Three question Chinese Abuse Assessment Screen (AAS)

Tiwari and colleagues compared three individual questions from the Chinese Abuse Assessment Screen with the three corresponding subscales for physical, emotional and sexual IPV from the 39 item Revised Chinese Conflict Tactics Scales (CTS2) in 100 pregnant and 157 non-pregnant Chinese women attending an antenatal clinic of a public hospital and a community centre in Hong Kong.[212]

Box 6: Three Chinese AAS questions

1. Within the last year, have you been physically hurt by someone?
2. Within the last year, have you been emotionally hurt by someone?
3. Within the last year, has anyone forced you to have sexual activities?

Quality Appraisal

The paper explicitly stated that the CTS2 was a "gold standard," [227] also having been validated using data from the first representative household study of spousal battering in Hong Kong.[228] It was unclear whether the reference standard was independent of the index tool. The Chinese AAS has been adapted and changed from the English AAS. It was impossible to judge it alongside the Chinese CTS2 for which a supporting reference was given which I was unable to access. Thus it was deemed unclear as to whether there was enough detail to replicate execution of the reference standard. This paper reports that the study is of Chinese women in Hong Kong. It is unclear whether Chinese is referring to the women's ethnicity, nationality or language spoken. There is some ethnic diversity in Hong Kong. Socioeconomic status was considered. The study population was less educated and poorer than the general population of Hong Kong.

3.2.2.2. Portuguese AAS's anchor question

In the second of the AAS papers, Reichenheim and colleagues, evaluated the test performance of the Portuguese AAS's anchor question on physical abuse during pregnancy against the 12 item physical violence scale of the modified Revised Portuguese conflict tactics scale (12 item mCTS2), in 748 post-natal Portuguese speaking women on the maternity wards of three public sector hospitals in Rio de Janeiro, Brazil.[213]

Box 7: Portuguese AAS's anchor question on physical IPV during pregnancy

Since you have been pregnant, have you been hit, slapped, kicked, or otherwise physically hurt by someone?

Quality Appraisal

The inclusion criteria were not explicitly stated in their Methods section though the exclusion criteria were clear. The paper stated that the translated Portuguese CTS2 was used as a standard, its content validity having been considered by evaluating its concept, item and semantic equivalences. Acceptable reliabilities were shown for each subscale, factor analysis identifying the underlying dimensions.[213] It should be noted that it was only the 12 item physical violence scale of the Portuguese CTS2 that was actually then used alone as the reference standard for physical IPV. It was unclear whether the reference standard was independent of the index tool. There was not enough detail in the paper to replicate the index tool. This was partly because the Portuguese versions were not in the paper but also because the paper stated in the Abstract that "…three anchor questions…..are the main focus of this article." However the evidence for validity was only collected using the one AAS anchor question on physical abuse during pregnancy. This paper reported that the study population was a Portuguese speaking population in Brazil. No further information was given about the ethnicity of this study population. Brazil is an ethnically diverse society. Socioeconomic data was collected showing that the study population was poorly educated and from low income families.

3.2.2.3. Ongoing Violence Assessment Tool (OVAT)

Ernst and colleagues investigated the four question Ongoing Violence Assessment Tool (OVAT) by testing it against the 30 question ISA in 306 women and men attending an Emergency Department in a US city, in a study population who were described as being 49% Caucasian, 20% Hispanic, 16% African-American and 15%

Asian or other.[214] OVAT's validity was not estimated according to these groups, hence one cannot determine if there were any differences in validity between them.

Box 8: Ongoing Violence Assessment Tool (OVAT) questions

1. Within the last month my partner has threatened me with a weapon
2. Within the last month my partner has beaten me so badly that I had to seek medical care
3. Within the last month my partner has had no respect for my feelings
4. Within the last month my partner has acted like he or she would like to kill me

OVAT- Questions 1, 2, & 4 are Yes No responses

Question 3 rated on a 5 point Likert scale - Never to Very Frequently

Quality Appraisal:

The paper explicitly stated that the ISA was the "gold standard" for detection of present ongoing IPV. It has already been noted above that the ISA only contains one question about sexual abuse. This reference standard (ISA) was not independent of the index tool (OVAT) as the OVAT was developed from questions in the ISA which had high predictive values but fewer Likert scale responses. It used the category of "Caucasian," when describing race.

3.2.2.4. HITS – two studies

Box 9: HITS (Hurt, Insult, Threaten, Scream) questions

1. How often does your partner physically hurt you?

2. How often does your partner insult you?

3. How often does your partner threaten you with harm?

4. How often does your partner scream or curse at you?

Answers to each item of HITS
☐Never ☐Rarely ☐Sometimes ☐Fairly often ☐Frequently

*Answers were summed to form an interval scale of the total HITS score, which could range from 4 to 20.

Chen and colleagues, evaluated the English and Spanish versions of the four question HITS, against different comparators – the English and Spanish versions of the 11 question ISA–Physical dimension (ISA-P, measures physical IPV) and the eight question WAST.[198] The study population were 202 English speaking and Spanish speaking Hispanic and non Hispanic women (72% Hispanic, 20% non-Hispanic White, 6% non-Hispanic Black, 1% non-Hispanic Other), attending an urban family practice in the US. They tried to compare the performance of the four HITS questions in the two different language groups.

Quality Appraisal
It is unclear whether the reference standard was acceptable. The situation is confusing due to the use of two comparators, in two languages which is in effect four reference standards in one paper. Certainly the evidence suggests that the 8 question English WAST is not an acceptable reference standard (see section 3.2.3.1.). The reference standard was not independent of the index tool. Most importantly not all the participants received the same reference standard. The study population's ethnicity profile reflected the practice population's of which 70% was also of Hispanic origin. There was also information on country of origin for the Hispanic women (with 39% Cuban / Cuban American, 35% Puerto Rican, 11% Dominican, 5% Mexican /

Mexican American, 10% other Latin American); and language spoken (44% of women completed interviews in Spanish). Socioeconomic status was measured as well as ethnicity. Hispanics and non-Hispanics were similar in all demographic characteristics. In contrast there were significant demographic differences between the two language groups - those who carried out the interview in English (would have included Hispanic women and non-Hispanic women) and those participants who completed the interview in Spanish (included only Hispanic women). The latter group tended to be older ($P < 0.001$), to have lower incomes ($P < 0.001$), to be married ($P < 0.001$), to have longer relationships ($P < 0.01$), and to be pregnant ($P < 0.05$). These significant differences persisted between Hispanic women who completed the interview in Spanish compared to Hispanic women who completed the interview in English. Spanish speaking Hispanic women were more likely to be Cuban and Cuban American ($P < 0.001$) (and less likely to be Puerto Rican and other Latin American) but tended to be older ($P < 0.001$), to have lower incomes ($P < 0.01$) and to be married ($P < 0.01$).[198]

The second paper examining the HITS, was by Sherin and colleagues.[215] The HITS was compared to a 15 item modified version of the CTS (15 item mCTS), consisting of the verbal and physical aggression items only. There were no ethnicity data describing the study population. The convergent validity correlation and internal consistency reliability were calculated during phase 1, in which 160 women were recruited from general practice. The diagnostic indices were generated during phase 2, a known group comparison of 99 self-identified survivors of IPV and 160 women from phase 1 (i.e. general patients visiting their physician). This would have included abused and non-abused women as suggested by the range of HITS scores generated in phase 1 (4-18).

Quality Appraisal

A known group comparison was used to calculate diagnostic accuracy. This study population was not representative of all women who attend general practice. Therefore there is no diagnostic accuracy evidence that HITS is able to identify women who had experienced abuse in a general clinical population. An acceptable reference standard was not used. This modified CTS had four reasoning items contained in the original CTS deducted from it on the basis that they were not directly

related to domestic violence. This left 15 physical and verbal violence items. Internal consistency reliability for the 15 item modified CTS in the same study population was 0.87. Apart from this evidence of validity based on internal structure there was no data presented external to the modified CTS to support that it could identify IPV. In addition to no ethnicity data there was also no socio-economic data about the study population.

3.2.2.5. Five non-graphic domestic violence (DV) questions

Zink and colleagues[216] compared five non-graphic domestic violence questions that can be used when children are present to the 39 item revised Conflict Tactics Scale[227] in 393 mothers recruited from primary care waiting rooms in Cincinnati, Ohio. 49% were white and 51% African American or other. Ethnicity was said to be potentially related to domestic violence status and hence included as a covariate in a logistic model evaluating the predictive ability of each question, by examining the areas under ROC curves.

Box 10: Five non-graphic domestic violence questions

1. How do you and your partner work out arguments?
2. In general how do you describe your relationship?
3. How is your partner treating you and the kids?
4. Do you feel safe in your current relationship?
5. Considering your current partners or friends or any past partners or friends, is there anyone who is making you feel unsafe now?

Likert format response scale, with 3 - 5 response options used for these 5 questions.

Quality Appraisal

Odds ratios adjusted for ethnicity (as well as age, education and income) were obtained from logistic regression. There were no significant differences in ROC areas between the five questions when logistic regressions were carried out. This implies that ethnicity is unlikely to affect the diagnostic accuracy of each question. This cannot be firmly concluded as the odds ratios presented were adjusted together for age, education, income and ethnicity. This study did in effect analyse ethnic differences in the five questions' validity whilst considering socio-economic factors.

3.2.2.6. Behavioural Risk Factor Surveillance Survey (BRFSS)

Bonomi and colleagues,[217] assessed the agreement between the Behavioural Risk Factor Surveillance Survey **(**BRFSS, five behavioural tactic abuse questions) with the Women's Experience with Battering scale (WEB, 10 impact questions), in a study population of 2,504 women accessed via telephone. They were 83% white and 4% Hispanic. This study was essentially looking at the validation evidence for the BRFSS questions in one ethnic group. There was no assessment of the validity of the BRFSS in any other ethnic group. Therefore no comparison can be made of BRFSS's validity in different ethnic groups.

Box 11: Behavioural Risk Factor Surveillance Survey (BRFSS) questions

1. Now I want to ask you about forced sex involving vaginal, oral, or anal penetration. Has an intimate partner ever forced you to participate in a sex act against your will?
2. Has an intimate partner ever threatened, coerced, or physically forced you into any sexual contact that did not result in intercourse or penetration?
3. Has an intimate partner ever hit, slapped, shoved, choked, kicked, shaken, or otherwise physically hurt you?
4. Have you ever been frightened for your safety, or that of your family or friends because of the anger or threats of an intimate partner?
5. Has an intimate partner ever put you down, or called you names repeatedly, or controlled your behavior?

Yes / no response scale used for these 5 questions.

Convergent validity was assessed as the BRFSS questions were directly compared to the WEB questions without labelling either as a reference standard. The numbers and percentages of women who were WEB positive / BRFSS positive, WEB negative / BRFSS negative, WEB negative / BRFSS positive and WEB positive / BRFSS negative were calculated without using correlation. It was possible to calculate diagnostic accuracy results from the data contained in this paper, for the different components of the BRFSS using the WEB, the longer tool as the comparator.

Quality Appraisal

It is not known whether the WEB is an acceptable reference standard (see Coker et al, 2001, section 3.2.2.7. and table 1).

3.2.2.7. Women's Experience with Battering Scale (WEB)

Coker and colleagues,[218] compared the ten question WEB to the 15 question modified ISA-Physical (15 item mISA-P),[226] in a study population of 1,152 participants who were recruited from two university family practice clinics in the US. 62% were African American and 38% white. Differences in these two groups were not analysed.

Box 12: Women's Experience with Battering Scale (WEB) questions

1. My partner made me feel unsafe even in my own home
2. I felt ashamed of the things my partner did to me
3. I tried not to rock the boat because I was afraid of what my partner might do
4. I felt like I was programmed to react a certain way
5. I felt like my partner kept me a prisoner
6. My partner could scare me without laying a hand on me
7. I hid the truth from others because I was afraid not to
8. I felt owned and controlled by my partner
9. My partner made me feel like I had no control over my life
10. My partner had a look that went straight through me and terrified me

WEB- Scored on a Likert scale of 1 (Strongly Disagree) to 6 (Strongly Agree)

Sum the responses for items 1 – 10. The range of scores is 10-60. Score of equal to or >20 indicates battering.

Association with external variables was measured by using correlation to assess the relationship between index scores and health indicators (number of physician visits in the last year and self-perceived poor mental health) – see table 3, on page 124.

Convergent validity was also assessed by using correlation coefficients between the WEB score and the 15 item mISA-P score without labelling either as a reference standard. From the numbers and percentages of women who were WEB+/mISA-P+, WEB-/mISA-P-, WEB-/mISA-P+ and WEB+/mISA-P- it was possible to calculate diagnostic accuracy results from the data contained in this paper, using the 15 item mISA-P, the longer tool as the comparator.

Quality Appraisal

The ISA-P is recognised as an acceptable reference standard to identify physical IPV. In the introduction to this paper, the authors stated that the ISA-P was being used. Later in the Method it was revealed that they were actually using a 15 item modified ISA-P as opposed to the original 25 item ISA-P assessing physical abuse. This 15 question modified ISA-P has a Cronbach's alpha of 0.93, suggesting high internal consistency. This high internal consistency gives no indication about whether the modified ISA-P questions are identifying IPV. Hence it is unclear whether the modified ISA-P is an acceptable reference standard. An area of bias that is not captured by the QUADAS is that the modified version of the ISA-P which assesses physical IPV is an inappropriate comparator for the WEB which is probably identifying emotional IPV as well as physical. Having an inappropriate reference standard takes precedence over other less important areas of bias.

I will now consider collectively these eight studies which reported validation paradigm methods but which also made use of diagnostic accuracy in order to decide what their results actually mean. This qualitative overview of their results also provides an opportunity to compare these two different types of methods generating validity evidence.

3.2.2.8. Overview of eight studies reporting validation paradigm methods with diagnostic accuracy

Tiwari and colleagues explicitly used methods from both paradigms.[212] Their use of diagnostic accuracy and Kappa coefficients was consistent with the data generated from the three Chinese AAS questions and Chinese CTS2 being categorical and dichotomous. However there was an incongruity in the results arising from these methods and their interpretation in that the sensitivity levels of the Chinese AAS questions (36 to 66%) were felt to be too low to be clinically useful whilst the kappa coefficients (0.56, 0.52 and 0.47) were interpreted as showing fair agreement between the Chinese AAS questions and the Chinese CTS2 by the authors.

Reichenheim and colleagues,[213] also used methods from both paradigms but by using the diagnostic accuracy model it treated the 12 question modified CTS2 data as being categorical whilst the point biserial correlation treated this CTS2 score as being dimensional. There was no discussion about whether it was reasonable or permissible to treat the CTS2 score as being both categorical and dimensional. There was a discrepancy again in the results from these different methods in that the point-biserial correlation of 0.68 was construed as being high, appearing to indicate that the AAS question was functioning well whilst the sensitivity and specificity indicated that two thirds of minor and one third of severe episodes of IPV were being missed. The PPVs were higher than the sensitivities with considerable differences in the pre- to post-test probabilities. This was less important than the poor content validation with neither sexual IPV nor emotional IPV being considered.

Ernst and colleagues,[214] use of diagnostic accuracy and Kappa statistic was consistent with the data generated from the OVAT and the ISA being categorical and dichotomous. Most of the diagnostic indices were high (except the PPV of 56%) whilst the kappa statistic of 0.58 was not interpreted by the authors. The PPV of 56% in an area of relatively high IPV prevalence (20% according to the ISA) is disappointing indicating a change from pre- to post-test probability of only 36%. In a lower prevalence area (for example primary care as opposed to an emergency department), the PPV is likely to be even lower.

Chen and colleagues,[198] used methods from both paradigms but did not state the statistical test used to calculate correlation. All the correlations were either more than or equal to 0.75. The sensitivities and specificities were also all above 86% but the PPV of the Spanish-HITS was only 45% with a resulting change from pre- to post-test probability of just 35%. Most importantly different comparators were used for the English and Spanish version of the HITS with no reasonable rationale for this disparity.

In Sherin and colleagues' study both correlation and diagnostic accuracy indices had consistently high results (r=0.85, sensitivity 96%, specificity 91%) but whilst correlation was calculated using a general practice population, the diagnostic indices were derived from a known group comparison.[215] Section 3.2.4.2. details the consequences of known group comparisons on study population heterogeneity and the impact on validity evidence, including diagnostic accuracy indices.

In these last four studies,[198, 213-215] I would suggest that the validity evidence based on relations to other variables is less important then the decreased validity evidence based on test content as sexual IPV is not explicitly assessed by either the Portuguese AAS single anchor question, OVAT or the HITS.

Zink et al[216] was the final study to use both the diagnostic accuracy method and from the validation paradigm, internal consistency reliability (see section 3.2.4.2.). In this study along with the last three studies, the validity evidence based on response processes is decreased by use of a Likert scale.[198, 215, 216, Ernst, 2004 #562] This potentially can cause problems for the respondent. More importantly, these questions cannot be used as part of a routine verbal history, taken in any clinical consultation.

Out of these six studies,[198, 212-216] five directly use diagnostic accuracy and criterion related correlation alongside each other.[198, 212-215] The first three studies suggest that the diagnostic accuracy data is more informative than methods correlating scores when assessing the validity of questions trying to identify IPV.[212-214] All five studies demonstrate that if the questions produce categorical data and a quality reference standard exists, diagnostic accuracy yields more clinically useful

information than just a single figure representing the correlation coefficient.[198, 212-215]

There were two further studies based on the validation paradigm which reported data allowing the computation of diagnostic accuracy results.[217, 218] In Bonomi et al's study,[217] the numbers and percentages of women who were BRFSS+/WEB+, BRFSS-/WEB-, BRFSS+ /WEB- and BRFSS-/WEB+, are not intuitive to interpret partly as there was no explicit theory describing how these two tools related to each other. Instead it was stated that they both identified "abuse" but no evidence was presented to support this. Generating diagnostic accuracy data provided one clearer interpretation of this data but forced the WEB to be a reference standard. It showed that the BRFSS for any kind of abuse had a low PPV; and a small difference between the pre- and post-test probabilities.

In Coker et al's study,[218] as in the Bonomi study, the numbers and percentages of women who were WEB+/mISA-P+, WEB-/mISA-P-, WEB+/mISA-P- and WEB-/mISA-P+ were presented but also initially difficult to construe. Calculating diagnostic accuracy helped to give more meaning to the data showing that WEB's PPV was only 52%. Clinically it is more meaningful to know that if a person answers positively that there is a 52% probability that she experiences physical IPV rather than a Cohen's kappa statistic of 60% indicating the agreement between two dichotomised measures.

I now consider the six studies that exclusively reported validation paradigm methods.

3.2.3. Six studies reporting exclusively validation paradigm methods

3.2.3.1. Woman Abuse Screening Tool (WAST) – three studies

Four papers evaluating the WAST were identified in this systematic review.[211, 219-221] Three of these studies followed the validation paradigm only, using a variety of different methods to generate evidence of validity but not diagnostic accuracy data.[219-221]

The first WAST study was conducted by Brown and colleagues.[219] In purposive samples of 24 abused and 24 non-abused women, the seven question WAST and the two question WAST-Short were compared to the Abuse Risk Inventory (ARI), in Canada. No ethnicity data was reported. The seven question WAST, unlike the eight question WAST, did not include the last question on sexual IPV. The two question WAST-Short were the two questions that women were most comfortable with from the seven question WAST.

Box 13: Woman Abuse Screening Tool-Short (WAST-short) questions

1. In general, how would you describe your relationship?
□ A lot of tension
□ Some tension
□ No tension
2. Do you and your partner work out arguments with:
□ Great difficulty?
□ Some difficulty?
□ No difficulty?

Quality Appraisal

The most important bias in this study was that the spectrum of patients was not representative of either the local population or of all women attending general practice. Instead two extreme groups of intentionally selected women were used, consisting of women from a local shelter for abused women and women accessed via the principal investigator's contacts. There was not enough detail to replicate the ARI and so to judge objectively whether the ARI was an acceptable reference standard. The lead author of the paper was unsuccessfully approached for a copy of the ARI. The actual reference supporting the ARI ("Yegidis BL. Abuse risk inventory manual. Palo Alto, Calif: Consulting Psychologist Press, 1989") was unobtainable. The ARI was said to have demonstrated reliability and validity in identifying women who are being abused by their partners though it was not specifically described as a standard criterion. The study population was small using only 24 abused women and 24 non-abused women. This study limitation was not highlighted by QUADAS.

The work of the first WAST study was developed by the second WAST study conducted by Brown et al.[220] Now an eight question WAST (included an extra question on sexual IPV) was compared to the Abuse Risk Inventory (ARI) in 307 women in a family practice setting (attending urban and rural family physician practices in South Western Ontario, Canada). The study population was homogeneous with 98% of it being white. All were English speaking. In effect this study was looking at the evidence of validity for the WAST questions in one specific group.

WAST's additional last question "Has your partner ever abused you sexually?" − is not clear in that the term "abuse" is quite technical and may not correspond to women's experiences of sexual IPV, for example being forced to have any kind of sexual activity or being raped. This decreases WAST's validity evidence based on response processes.

Quality Appraisal

There was not enough detail to replicate the ARI and so to judge objectively whether the ARI was an acceptable reference standard (see above). In this paper, the ARI's

role was not of a standard criterion but the theoretical relationship linking the ARI to the WAST was not explicitly discussed.

Brown repeated the study in a Francophone community, again using a convenience sample of 25 abused women residing in two women's shelters and 21 non-abused women, in Ontario and Quebec, using a French version of the WAST and the WAST-Short.[221] No further information was given on the ethnicity of the study population.

Quality Appraisal

The spectrum of patients was not representative being a known group comparison. This study, in common with the earlier study from 1996, had a very small study population – with 25 abused and 21 not abused. The study population was a Francophone community in Ontario and Quebec. No further ethnicity details were provided. The authors stated that the abused and not abused women were demographically similar. This was surprising given that 9% of the abused women were employed and 92% of the non-abused; 32% of the abused women were married and 81% of the non-abused women (see table 2, on page 121).

3.2.3.2. WAST-Short and HITS

Most recently Chen and colleagues, in 2007, have validated both the English WAST – Short (two questions) and the English HITS (four questions) by comparison to the eight question WAST.[222] This was in a study population of 523 minority women, predominantly African American (71%) attending four urban family medicine practices. Ethnicity was not used as a study variable to assess validity evidence between groups. Convergent validity was assessed using correlations of the WAST – Short and the HITS with the eight item WAST.

Quality Appraisal

The eight question WAST is not an acceptable reference standard – as shown recently in a study.[211] However in this paper its role was not that of a standard criterion. The theoretical relationship linking the eight question WAST to either the Short-WAST or the HITS, was not explicitly described. The eight question WAST was not independent of the index tool, the WAST-Short, in that both of them were developed together.[219-221] This causes incorporation bias as the comparator (eight question WAST) in effect includes the WAST-Short, possibly increasing multicolinearity. This probably explains the high correlation of the WAST–Short score with the WAST total score (0.81, p<.001).

3.2.3.3. Perinatal Self-Administered Inventory (PSAI)

Sagrestano and colleagues compared the 2 questions on IPV within the Perinatal Self-Administered Inventory (PSAI) to the seven verbal aggression questions and nine physical violence questions of the CTS in 166 women in antenatal clinics [223] of whom 48% were African American, 46% Hispanic, 6% white or other. 25% (n=42) completed the interview in Spanish. There was no analysis to examine the difference between either the different ethnic groups, or the two different language groups.

Box 14: Two Perinatal Self-Administered Inventory questions

First question: Are you experiencing severe conflicts with anyone in your home?

Second question: Are you suffering mental or physical violence abuse now?

Both the PSAI questions were complex impacting on the response processes and the consequential validity evidence. The first question ("are you experiencing severe conflicts with anyone in your home?") is not only asking whether there is conflict (what is conflict?) at home, but whether it is severe (how should severity be graded?)

and "…anyone at home" is non-specific. The second question enquires about abuse rather than a specific act, lumping together mental and physical violence and asks whether it is happening right now (what does "right now" mean? "Right now" I am being interviewed and not being abused or does "right now" mean today or this week or this month etc.). This timescale is very different from that in the CTS ("in the last year"). Therefore it is not surprising that it neither correlated to either the verbal or physical abuse sub-scales of the CTS.

Quality Appraisal

The time period was not short enough between administered tools with an average time lag of three weeks mentioned. The CTS subscales of verbal aggression and physical violence are not known to be acceptable reference standards. Neither were they independent of the index tool. However in this paper their role was not that of standard criterions but the theoretical relationship linking the PSAI questions and the CTS was not explicitly described. The study population was diverse with 25% (n=42) completing the interview in Spanish.

3.2.3.4. Three question English AAS

McFarlane and colleagues[224] tested three questions from the AAS (numbers 2, 3 & 4 – see Box 15, on page 159) against the 30 item ISA, using correlation to assess convergent validity, in 691 pregnant women in Houston, Texas and Baltimore in the US. The questions were offered in English and Spanish. The number of women who completed the questions in Spanish was not stated. There was no comparison of AAS's validity between specific language or ethnic groups in this study.

Box 15: Three questions from the AAS

2. Within the last year, have you been hit, slapped, kicked or otherwise physically hurt by someone?
3. Since you've been pregnant, have you been hit, slapped, kicked, or otherwise physically hurt by someone?
4. Within the last year, has anyone forced you to have sexual activities?

Quality Appraisal

The results were un-interpretable with no data presented to support that those positively screened for IPV with the AAS were more likely to have a significantly higher score on the ISA. The authors still concluded that the AAS questions were valid and specific in identifying abuse.

3.2.4. Overview of correlation measures

Correlation measures appeared in the six studies above that exclusively reported validation paradigm methods along with some of the earlier studies which used intersecting methods. Correlation measures have been used in these studies to measure validity evidence based on relations to other variables (for example criterion related validity and convergent validity) as well as validity evidence based on internal structure (i.e. internal consistency reliability). The analysis of these correlation measures collectively is presented below.

3.2.4.1. Validity evidence based on relations to other variables

See table 3 for a summary of studies that used correlation measures to assess criterion correlation validity, (i.e. used a reference standard), convergent validity and the

association between index scores with external variables. This table includes a précis of each paper's interpretation of this data.

Out of the 20 studies included in the systematic review, five studies used correlation to establish criterion related validity.[198, 212-215] Two of these studies used the kappa coefficient to calculate the criterion related validity,[212, 214] one used biserial correlation,[213] and two did not state the statistical method used.[198, 215] The criterion related validity coefficients ranged from 0.47 to 0.85. The lowest criterion related validity coefficients was seen in the Tiwari study - a kappa coefficient of 0.47 for sexual IPV.[212]

Eight other studies were association studies, containing 21 estimates of convergent validity for nine sets of questions trying to identify IPV.[217-224] The majority of these were correlation coefficients though in two cases convergent validity was expressed simply using the numbers and percentages of women in overlapping groups.[217, 218] The statistical method used to calculate the correlation coefficient was most commonly not stated (nine instances). When the statistical method was stated Pearson's correlation coefficient was used for two estimates,[218, 220] Cohen's kappa coefficient for one estimate[218] and Spearman correlation for one.[219] The convergent validity correlation coefficients ranged from 0.03 to 0.96.

None of the eight studies examining convergent validity contained explicit information on the theories about IPV linking different sets of questions. This is a central issue to interpreting the meaning of a correlation coefficient. Instead there often appeared to be the assumption that the higher the correlation coefficient, the better the index set of questions. For example, Coker et al[218] in their introduction considered that the WEB identified battering (related to loss of power and control) whilst the ISA assessed episodic physical assaults. There was recognition that both were conceptually and empirically distinct but the method still measured the agreement between the two measures using correlation.

I now examine the studies that used correlation measures to measure internal consistency reliability.

3.2.4.2. Validity evidence based on internal structure

See table 4 for a summary of studies that used correlation measures to assess internal consistency including each paper's interpretation of the values. Eight studies contained 11 estimates of internal consistency for eight sets of questions trying to identify IPV.[198, 214-216, 219-222] Ten of these measures were Cronbach's alpha whilst one was a corrected item-total correlation. The Cronbach alphas ranged from 0.46 (interpreted as being mediocre) to 0.95 (interpreted as being good).

None of the papers contained any explicit discussion about whether the sets of questions trying to identify IPV

i. were operating over one dimension (uni-dimensional) or many dimensions of IPV (multi-dimensional) or
ii. comprised categorical or dimensional data

These are central principles that should be considered prior to calculating internal consistency measures.

Out of the eight sets of questions identifying IPV, five operate over more than one dimension. Both the English HITS (and presumably the Spanish HITS), the OVAT and the seven question WAST contain questions about physical IPV and emotional IPV. Hence they are not uni-dimensional but instead operate across two different dimensions. The eight question WAST contains questions about physical, emotional and sexual IPV, covering three dimensions of IPV. It is inappropriate to measure internal consistency for these question sets as done by seven of these studies.[198, 214, 215, 219, 220, 221, 222] Zink's five non-graphic questions are said to cover the major domains of domestic violence, including personal safety, the treatment of children as well as containing the 2 WAST-Short questions.[216] It therefore also seems unlikely that these five questions will be operating over just one dimension for which it is appropriate to apply an internal consistency reliability measure. All the sets of questions were trying to categorise women into those experiencing IPV and those that were not. Hence overall the evidence points to internal consistency measures not being applied appropriately in the majority of these studies.

The English WAST-Short and presumably the French WAST-Short are uni-dimensional. Both had high Cronbach alphas of 0.8 and 0.95 respectively, interpreted as being good.[221, 222] However it is important to note that the Cronbach alpha (in common with all correlation coefficients), is not just a reflection of internal consistency (or how strong the relationship is between 2 measures) but also the study population's heterogeneity, (see section 1.3.6.4.)

This is demonstrated by this systematic review which shows that in the two known group comparisons, used to generate not only internal consistency reliability (Cronbach's alphas) but also convergent validity correlation coefficients, the results were the highest seen.[219, 221] The Cronbach alpha for the 7 question WAST was 0.95,[219] whilst the Cronbach alpha for the French WAST was 0.95.[221] For the total 7 question WAST, r=0.96, for individual WAST questions r=0.80 to 0.85.[219] For the total 8 question French WAST, r=0.96, for individual WAST questions r=0.75 to 0.93.[221] These results most likely reflect that both these two known group comparisons contained two extreme populations of abused and non-abused women with the widest score ranges and greatest study population heterogeneity. It does not necessarily indicate that these questions were most highly correlated to each other (internal consistency reliability) or that they were the most highly correlated to other instruments (convergent validity).

The heterogeneity of the study population was supported in the Brown 1996 study by the significant differences found in all socioeconomic indicators (employment status, income and education), age and percentage married between the two groups of abused and not abused women.

These two known group comparisons and a further known group comparison was also used to generate diagnostic accuracy indices.[215, 219, 221] These figures were also artificially inflated due to the increased study population heterogeneity. This makes their values incomparable to those derived in studies using participants representative of patients attending general practice (see Table 1, on page 110).

3.2.5. One study using neither research paradigm

Connelly and colleagues[225] tested a single question which was part of a hospital admission protocol against the 9 question physical subscale of the CTS, the comparator, in 436 high risk post partum mothers. The study population was 40% Hispanic, 23% African-American, 27% Caucasian and 9% Asian, Pacific Islander, Native American and other. The validation of the single question was not compared between these groups.

Box 16: One question from hospital admission protocol

"Are you in a relationship in which you have been threatened, scared or hurt by someone?

If yes, whom?"

Apart from presenting the prevalence of IPV according to the CTS and the percentage of the sample that were threatened, scared or hurt, the relationship between the two was not analysed.

Quality Appraisal

The study population spectrum was not representative of all patients instead being high risk post-partum mothers in whom the risk of moderate to severe violence is thought to be greatest.[229, 230] An acceptable reference standard was not used. The time period was not short enough between administered tools.

I have now presented all the systematic review results. I now present the results of my secondary data analysis.

3.3. Secondary data analysis results

The chief findings of my secondary data analysis results are presented in tables showing diagnostic accuracy indices with 95% confidence intervals of HARK at different cut off scores for the south Asian, African-Caribbean and white groups (see tables 7, 8 and 9). There is also a receiver operator characteristic curve for each ethnic group (see figures 6, 7 and 8). Figure 9 compares the three receiver operator characteristic curves generated by the three groups. For a complete record of my secondary data analysis, see Appendix F. My commentary focuses on what potentially may be clinically important results.

Table 7: The sensitivity, specificity, PPV, NPV, LR (with 95% confidence intervals), post-test odds and pre- to post-test probability of IPV at different HARK cut off scores, in the African-Caribbean groups (N = 59).

Hark cut off scores	% of study sample	Sensitivity with 95% C.I.	Specificity with 95% C.I.	Positive predictive value with 95% C.I.	Negative predictive value with 95% C.I.	Likelihood ratio with 95% C.I.	Post-test odds	Change from pre- to post-test probability of IPV
= 4	2%	5% (0% to 28%)	100% (89% to 100%)	100% (5% to 100%)	69% (55% to 80%)	Undefined	Undefined	68% (62% to 74%)
≥ 3	10%	32% (13% to 56%)	100% (89% to 100%)	100% (52% to 100%)	75.5% (61% to 86%)	Undefined	Undefined	68% (62% to 74%)
≥ 2	15%	47% (25% to 70%)	100% (89% to 100%)	100% (63% to 100%)	80% (66% to 89%)	Undefined	Undefined	68% (62% to 74%)
≥ 1	34%	89.5% (65% to 98%)	92.5% (78% to 98%)	85% (61% to 96%)	95% (81% to 99%)	12 (4 to 36)	6	53% (45% to 60%)
≥ 0	100%	100% (79% to 100%)	0% (0% to 11%)	32% (21% to 46%)	Undefined	1	0.5	0% (-8% to 8%)

When the specificity is 100%, the likelihood ratio and post test odds are undefined. Confidence intervals for likelihood ratios are approximate since they were calculated by the delta method which is less reliable when some cell sizes are small (Armitage P, Matthews JNS, Berry G. *Statistical methods in medical research.* Oxford; Blackwell, 1994).

Table 8: The sensitivity, specificity, PPV, NPV, LR (with 95% confidence intervals), post-test odds and pre- to post-test probability of IPV at different HARK cut off scores, in the south Asian groups (N = 48).

Hark cut off scores	% of study sample	Sensitivity with 95% C.I.	Specificity with 95% C.I.	Positive predictive value with 95% C.I.	Negative predictive value with 95% C.I.	Likelihood ratio with 95% C.I.	Post-test odds	Change from pre- to post-test probability of IPV
= 4	0%	0% (0% to 30%)	100% (88% to 100%)	Undefined	75% (60% to 86%)	Undefined	Undefined	-
≥ 3	4%	17% (3% to 49%)	100% (88% to 100%)	100% (20% to 100%)	78% (63% to 88%)	Undefined	Undefined	75% (69% to 80%)
≥ 2	12.5%	42% (16% to 71%)	97% (84% to 100%)	83% (36% to 99%)	83% (68% to 92%)	15 (2 to 116)	5	58% (50% to 65%)
≥ 1	21%	75% (43% to 93%)	97% (84% to 100%)	90% (54% to 99%)	92% (77% to 98%)	27 (4 to 192)	9	65% (58% to 71%)
≥ 0	100%	100% (70% to 100%)	0% (0% to 12%)	25% (14% to 40%)	Undefined	1	0.3	0% (-8% to 8%)

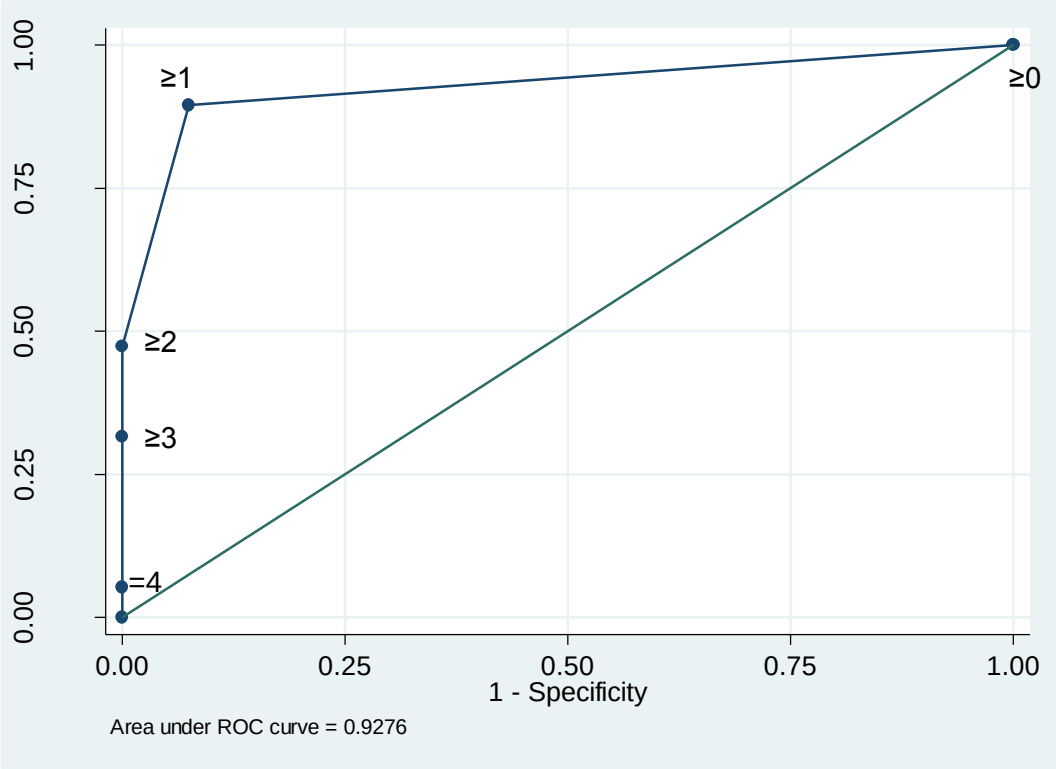
When the specificity is 100%, the likelihood ratio and post test odds are undefined. Confidence intervals for likelihood ratios are approximate since they were calculated by the delta method which is less reliable when some cell sizes are small (Armitage P, Matthews JNS, Berry G. *Statistical methods in medical research.* Oxford; Blackwell, 1994).

Table 9: The sensitivity, specificity, PPV, NPV, LR (with 95% confidence intervals), post-test odds and pre- to post-test probability of IPV at different HARK cut off scores, in the white groups (N = 112).

Hark cut off scores	% of study sample	Sensitivity with 95% C.I.	Specificity with 95% C.I.	Positive predictive value with 95% C.I.	Negative predictive value with 95% C.I.	Likelihood ratio with 95% C.I.	Post-test odds	Change from pre- to post-test probability of IPV
= 4	1%	5% (0% to 27%)	100% (95% to 100%)	100% (52 to 100%)	83% (74% to 89%)	Undefined	Undefined	77% (71% to 82%)
≥ 3	5%	30% (13% to 54%)	100% (95% to 100%)	100% (52% to 100%)	87% (79% to 92%)	Undefined	Undefined	77% (71% to 82%)
≥ 2	13%	65% (41% to 84%)	98% (92% to 100%)	87% (58% to 98%)	93% (85% to 97%)	30 (7 to 122)	6.5	64% (57% to 70%)
≥ 1	17%	75% (51% to 90%)	96% (89% to 99%)	79% (54% to 93%)	95% (87% to 98%)	17 (6 to 46)	4	56% (48% to 63%)
≥ 0	100%	100% (80% to 100%)	0% (0% to 5%)	18% (11% to 26%)	Undefined	1	0.2	- 5%

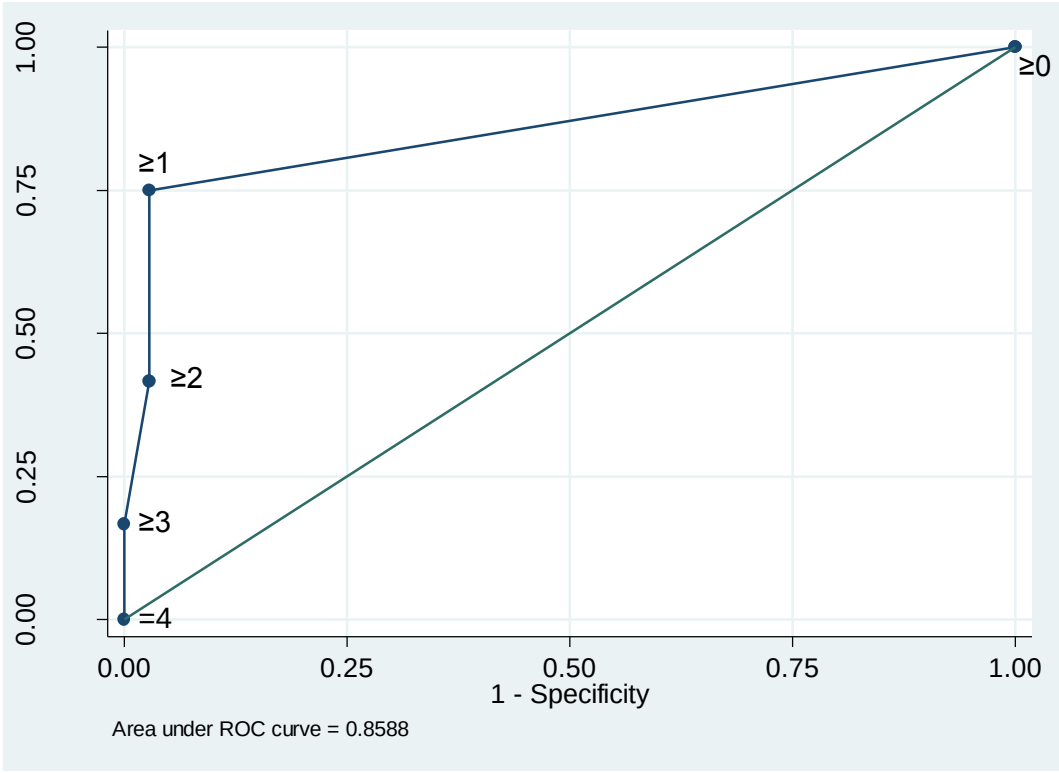
When the specificity is 100%, the likelihood ratio and post test odds are undefined. Confidence intervals for likelihood ratios are approximate since they were calculated by the delta method which is less reliable when some cell sizes are small (Armitage P, Matthews JNS, Berry G. *Statistical methods in medical research.* Oxford; Blackwell, 1994).

Figure 8: Receiver operator characteristic curve for the African-Caribbean groups, showing sensitivity of different HARK scores verses 1 - specificity



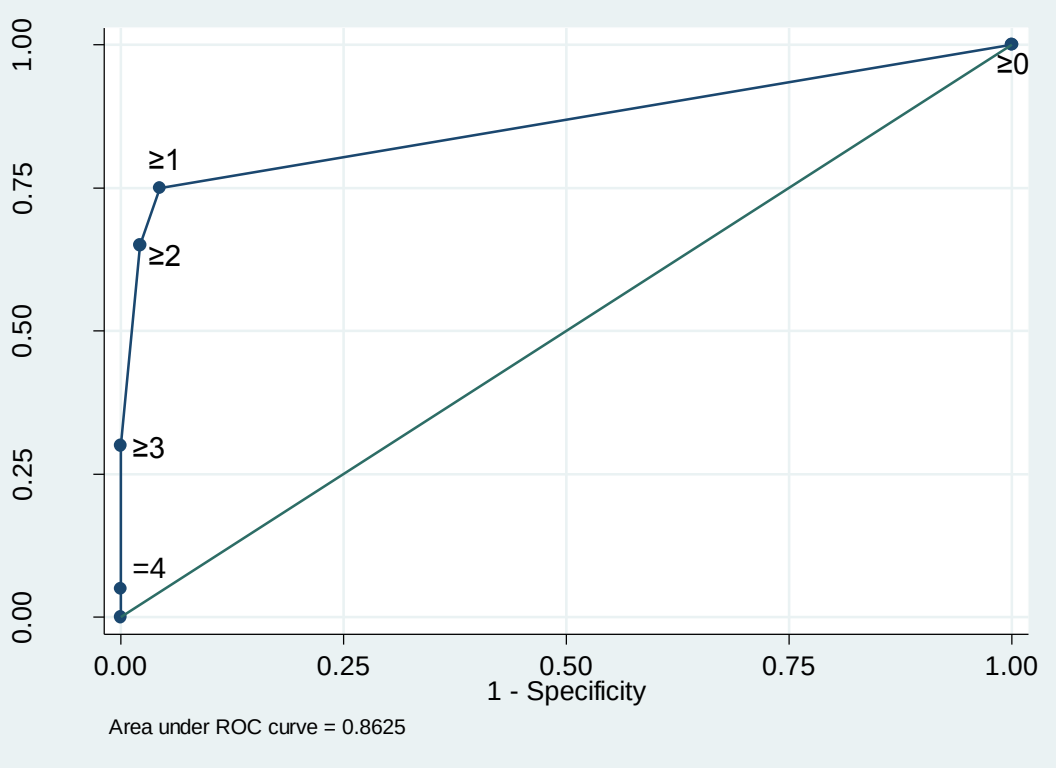
95% confidence interval for area under ROC curve: 0.85 to 1.00

Figure 9: Receiver operator characteristic curve for the south Asian groups, showing sensitivity of different HARK scores verses 1 - specificity



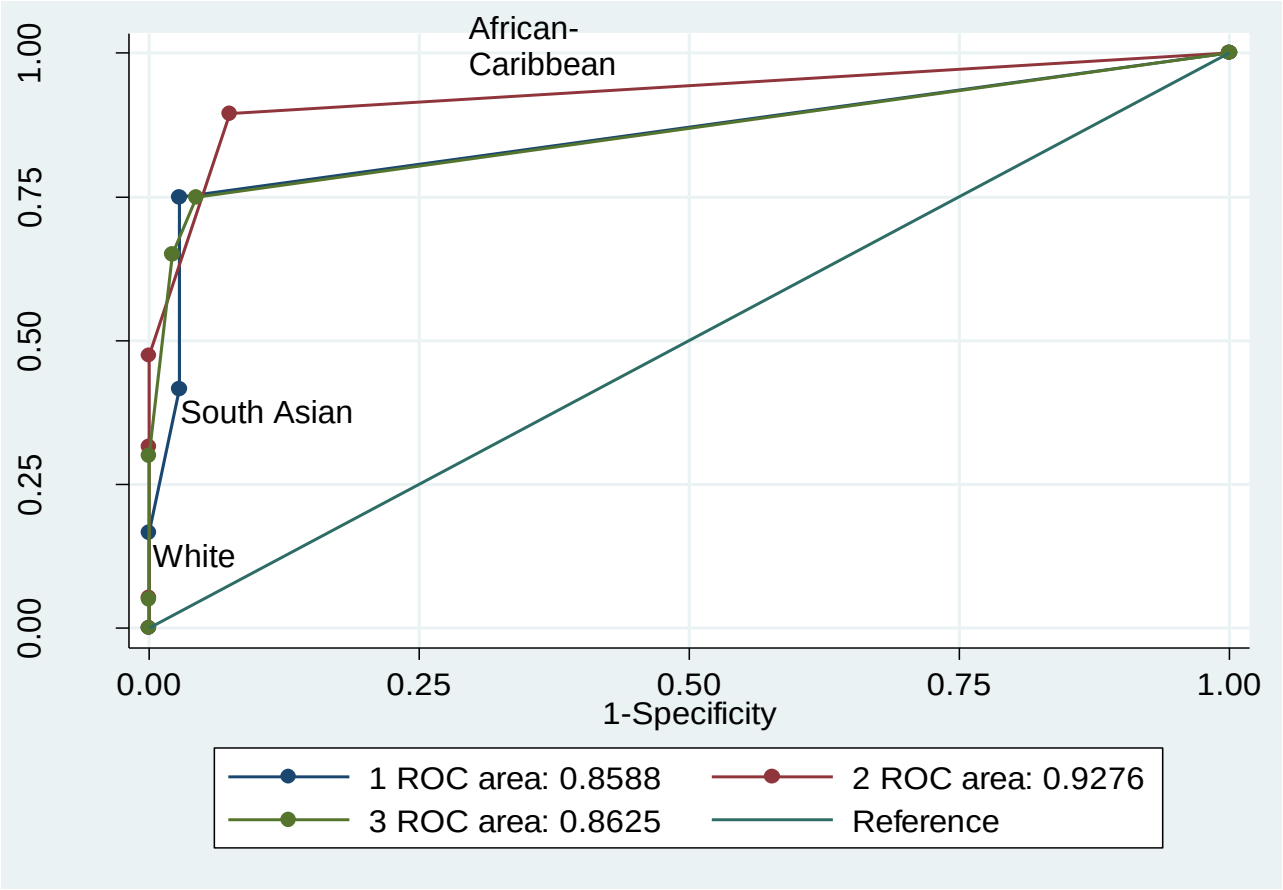
95% confidence interval for area under ROC curve: 0.73 to 0.99

Figure 10: Receiver operator characteristic curve for the white groups, showing sensitivity of different HARK scores verses 1 - specificity



95% confidence interval for area under ROC curve: 0.76 to 0.96

Figure 11: Comparing the three receiver operator characteristic curves in the African-Caribbean, south Asian and white groups



3.3.1. Commentary

Tables 7, 8 and 9 (on pages 166 to 168) consistently show wide and overlapping 95% confidence intervals suggesting that this study was underpowered to detect any statistically significant ethnic differences in the ability of HARK to identify IPV. Therefore my commentary rather than focussing on the statistical significance of these results will focus on what potentially may be clinically important results. This includes the ethnic similarities seen in the data as well as some of the differences seen in the validity evidence for the HARK questions' ability at identifying IPV.

For all three groups, African-Caribbean, south Asian and white, the receiver operator characteristic curves (see figures 7, 8 and 9, on pages 169 to 171) clearly demonstrate that a HARK score of ≥1 is the optimal cut off for identifying IPV, as it is for the entire population.[200] This cut off maximises the true positives whilst minimising the false positives, in each group. Figure 10 (on page 172) shows that there was no significant variation in the areas under the three ROC curves for the three groups, (also see Appendix E, page 276).

The diagnostic indices generated using the HARK cut off of ≥ 1 were at a high level, for the African-Caribbean, south Asian and white groups. Most importantly in all three groups, using the HARK questions resulted in wide differences in the pre- to post-test probabilities of IPV (53%, 65% and 56% respectively). There appeared to be no statistically significant differences in the diagnostic indices between these three groups. This was not unexpected, however, because the study did not have the power to detect differences between the ethnic groups of the orders that are to be expected. The simple scoring system and good content validation of the HARK questions was equally applicable to all three ethnic groups as all the study participants were English speaking, completing the HARK questions in English.

The kick question for identifying IPV also appeared to operate in the same way in the three groups in that it had a PPV and specificity of 100% and consequently undefined

LRs and PTOs (approaching infinity) in all three groups. The pre- to post-test probability of IPV detected using the kick question was 77%. The kick question for identifying physical IPV also had 100% PPV and specificity with undefined LRs and PTOs again in all three groups. The change from pre- to post-test probability of physical IPV detected whilst using the kick question to identify physical IPV was 82%.

In contrast, the afraid question for identifying IPV and for identifying emotional IPV only had 100% PPVs and specificities with consequently undefined LRs and PTOs (approaching infinity) in the African-Caribbean groups. When using the afraid question for identifying IPV in the African-Caribbean groups, there was also a wider difference in the pre- to post-test probability of IPV (100 – 32 = 68%) than when actually using all four HARK questions (53%). When using the afraid question for identifying emotional IPV in the African-Caribbean groups, there was also a wide difference in the pre- to post-test probabilities of emotional IPV (100 – 25 = 75%).

Only three women answered "yes" to the HARK "rape" question. Therefore it was decided not to examine sexual IPV (or the CAS dimension of severe combined abuse which includes sexual IPV) as this number was too small for any meaningful analysis.

I have now presented all my results. In the next chapter, I will discuss these results.

Chapter 4: Discussion

4.1. Overview

In this chapter I will be discussing my results and their implications. I will first summarise the answer to my principal research question and then my related secondary research question using the results of the systematic review and my data analysis. I will then consider why these findings are important and the potential impact that they have for clinical practice. Next there is discussion about my quality appraisal of methodology and my quality appraisal of the use of ethnicity data in the 20 primary studies included in the systematic review and my secondary data analysis. Following on from this, in section 4.6., I will consider the limitations of QUADAS as a quality appraisal tool. I compare QUADAS to the *Sta*ndards for *R*eporting of *D*iagnostic Accuracy (STARD) statement which is also concerned with the quality of the methods of diagnostic accuracy studies.

In section 4.7., I will consider the strengths of my thesis which builds on previous work into identifying IPV. Many of the strengths centre on my bringing of psychometric principles to bear on the subject of IPV identification and applying the evidence base for clinical diagnosis to IPV identification. I used the Standards for educational and psychological testing as the basis for my comprehensive categorisation of evidence of validity and throughout my thesis employed contemporary terminology which these Standards promoted. Additionally, my focus on first principles has resulted in my drawing attention to the most important function of index questions, i.e. being able to change the pre-test probability of IPV to the post-test probability of IPV by the greatest percentage. Following this discussion, I will consider the limitations of my work. These relate to financial constraints and largely failing to study language as an integral component of ethnicity.

4.2. What is the evidence for the validity of questions trying to identify IPV in specific ethnic groups?

The systematic review showed that in six studies the study population in effect consisted of one predominant group, allowing the evidence for the validity of questions used in specific ethnic groups to be assessed: African-Americans (see section 3.2.1.2), Chinese women in Hong Kong (see section 3.2.2.1.), white Americans (see section 3.2.2.6.), white English speaking Canadians, French speaking Canadians (see section 3.2.3.1.) and English speaking Canadian born women (see section 3.2.1.3.).

The evidence for the validity of the STaT questions to identify IPV in an African-American population shows that they could not be used to identify IPV in brief clinical consultations. This is verified by the STaT not identifying sexual IPV and its low PPV (all < 48%) regardless of the STaT cut off point used. In a population with a high prevalence of IPV (33% according to the ISA), having PPVs of this magnitude is too low to be of practical use to a clinician. This is reflected by the change from pre-test probability to post-test probability of IPV being only 15% at most.

The evidence for the validity of the two individual questions from the Chinese AAS to identify physical and sexual IPV in Chinese women in Hong Kong shows that they also could not be used in clinical practice. This is confirmed by sensitivities of 45% and 36% suggesting one could not be confident that either physical IPV or sexual IPV (respectively) is being identified. The PPVs were higher than the sensitivities with considerable change from the pre- to post-test probability, for physical and sexual IPV. This means that individuals that test positive on these Chinese AAS questions are likely to have IPV but many other women with IPV will not test positive at all with these questions. Sackett reasoned that predictive values were more important than either sensitivity or specificity in identification (see Background, section 1.3.6.2.1.1.). I would contend that in IPV identification, ideally questions should have a high PPV combined with a high sensitivity so that not only women who test positive are likely to have IPV

but also so that most women with IPV do test positive. These two Chinese AAS questions both have a high NPV combined with a high specificity so that not only are women who test negative unlikely to have IPV but most women who do not experience either physical or sexual IPV do test negative. For IPV identification, questions that have high predictive values combined with high sensitivities and specificities are better than those that only have high predictive values.

The evidence for the validity of the BRFSS questions to identify IPV in a white American population is inconclusive as though the BRFSS appears to have a moderate level of diagnostic accuracy for any kind of abuse, this study needs to be repeated using a verified reference standard, instead of the WEB.

In a white English speaking Canadian family practice population the eight question WAST's correlation with the ARI of r=0.69 is impossible to interpret as there is not enough detail to replicate execution of the ARI and internal consistency reliability (Cronbach's coefficient 0.75) should not have been applied to this multi-dimensional scale.

In a French speaking Canadian population, the French eight question WAST's correlation with the ARI of r = 0.96 is also impossible to interpret as there is not enough detail to replicate execution of the ARI. The high value of r foremost represents the heterogeneity of extreme groups generated by using a known group comparison, as opposed to the closeness of the relationship between the two measures. This second point is equally true of the internal consistency coefficient alpha of 0.95 for the French WAST and for the sensitivity / specificity values given for the two question French WAST-Short.

A study in a predominantly Canadian born English speaking population showed that the eight question WAST and the PVS cannot identify IPV in this population in brief clinical consultations. The low sensitivities for both the WAST and the PVS means that both were not identifying women who were identified with IPV on the CAS. This was accompanied by low PPVs.

Overall, with regard to the validity of questions trying to identify IPV in specific ethnic groups (including the white groups) analysed in the systematic review, there is insufficient evidence to justify their use in clinical practice.

My secondary data analysis showed that in self-classified UK census categories of south Asian, African-Caribbean and white groups, the four HARK questions whilst using a cut off of ≥1 were able to identify IPV as shown by high diagnostic accuracy indices (predictive values as well as sensitivity and specificity) and the four HARK questions produced a substantial difference between the pre- to post-test probabilities of IPV (see tables 7, 8 and 9, on pages 166 to 168).

The secondary data analysis, unlike the systematic review, provides tentative evidence for the validity of questions trying to identify IPV in some specific ethnic groups (i.e. self-classified national census categories of Asian, African-Caribbean and white groups).

I now summarise the answer to my second research question.

4.3. Does the evidence for the validity of questions trying to identify IPV vary between different ethnic groups?

None of the six studies in which the study population consisted of one ethnic group were repeated in another ethnic group. Therefore comparisons between groups from different studies were not possible. Four further studies did try to analyse the differences between ethnic groups within the same study. Two of these studies examined the ethnic differences in the validity evidence for questions trying to identify IPV.[198, 216]

Chen and colleagues looked at the differences in two ethnic groups - English speaking and Spanish speaking Hispanic and non Hispanic women.[198] Their use of a different comparator in each of the two groups means that a direct comparison cannot be made. The study still concluded that there was a difference in the use of HITS in the two groups (cut-off score for the Spanish HITS was half of the cut-off score for the English HITS) which they attributed to culture. No data were measured reflecting cultural attributes apart from the language difference itself. The results did show that there were clear socio-economic differences between the English and Spanish speaking groups which the authors thought supported the cultural differences between the two language groups. However Hispanics and non-Hispanics were similar in all socio-economic characteristics, though are likely to have cultural differences. It is misleading to conflate culture with socio-economic status.

Zink and colleagues found that ethnicity (white versus African American) combined with age, education and income did not significantly affect the diagnostic accuracy of five non-graphic questions used to identify IPV.[216] This finding is consistent with the five questions being equally valid in white and African American populations. However the sensitivities and PPVs for these questions were small, they did not demonstrate internal consistency reliability, a Likert type scale (see section 1.3.6.5.) was used and the content

validation of these questions indicated that they did not explicitly ask about physical or sexual IPV. This last point is to be expected as the questions were trying to be non-graphic. Poor evidence of validity based on the questions' content could be overcome by high diagnostic indices. However when this is not achieved the failure of the contents to have validity becomes paramount. Overall this validity evidence does not support the use of these questions in white or African American populations.

My secondary data analysis showed that there were no statistically significant ethnic differences in the ability of the HARK questions to identify IPV or its dimensions of physical or emotional IPV. This included when the four HARK questions were used together to identify IPV and when the HARK questions were used individually to identify IPV and its dimensions, i.e. the "kick" question to identify physical IPV, the "humiliation" question to identify emotional IPV and the "afraid" question to identify emotional IPV. However the analysis is underpowered to detect differences of the magnitude that may occur between the ethnic groups.

The "afraid" question seemed to be more valid in the African-Caribbean groups than the south Asian or white groups at identifying both IPV and emotional IPV (see section 3.3.1.) in that its PPV for both was 100% with wide differences in the pre- to post-test probabilities. However the differences in these and all diagnostic indices for the afraid question between the three groups were not statistically significant (illustrated by wide overlapping confidence intervals). Hence checking for confounding by socio-economic status was not required. It is plausible that the very high PPV of the "afraid" question in the African-Caribbean groups simply reflects higher IPV prevalences which are manifests of possibly lower socio-economic statuses in the African-Caribbean groups. Alternatively this could be a potentially clinically important ethnic difference (see section 4.4.).

Overall my secondary data analysis concurs with the primary studies included in the systematic review: there is nothing to suggest that the evidence for the validity of questions trying to identify IPV varies between ethnic groups.

I now consider why the answers to my two research questions are important by putting them in the context of previous reviews and examining their potential impact for clinical practice.

4.4. Findings in context of other reviews and clinical practice

These are important results as firstly they are novel and very different from the results of previous systematic reviews looking at questions to identify IPV. Secondly, these results are also potentially clinically important as they may impact on clinical practice. I will expand on both of these areas below.

In previous systematic reviews a number of the questions that I have judged to not be useful for clinical practice have been commended. I think that these distinct conclusions arise as a result of previous reviewers taking numerical results at face value rather than as in this study in which I have worked from first principles (as described in the Background chapter) and examined the process by which these figures have been generated. This has allowed me to judge the legitimacy of data. I will now compare the results of four previous systematic reviews[7, 50, 231, 232] with the results of my systematic review.

The Canadian Task Force on Preventive Health Care in 2003, after having conducted a systematic review concluded that the WAST had "…acceptable psychometric properties …for primary care settings…".[231] This appears to have been based on evidence from three studies.[219-221] I have interpreted the evidence for the WAST questions from these studies as not being particularly compelling: internal consistency reliability measures have been used inappropriately and convergent validity was un-interpretable as the ARI was not accessible (see section 3.2.3.1.). Additionally values for internal consistency reliability and convergent validity are amplified having been derived from known group comparisons.[219, 221] It was on the basis of this Canadian Task Force systematic review that MacMillan and colleagues,[211] from the McMaster Violence Against Women Research Group decided to use the eight question WAST as index questions in their validation study. Their diagnostic accuracy data showed that WAST's sensitivity was 47%, PPV 55% and that the WAST questions changed the pre-test probability to post-test probability of IPV by 45%. The low values of these diagnostic

indices were attributed to errors associated with the reference standard, the CAS, though this was also recognised as the most comprehensive measure of IPV. This interpretation is more debatable in the context of the HARK questions achieving a sensitivity of 81%, a PPV of 83% and a change from pre- to post-test probability of 60% whilst using the same comparator (CAS) as the MacMillan study. I would argue that the diagnostic accuracy data from the MacMillan study confirms that the WAST cannot accurately identify IPV. Despite the poor evidence supporting the WAST, it continues to be used.[233]

The U. S. Preventive Services Task Force's systematic review from 2004 supported the use of the HITS and the WEB,[50] unlike my systematic review. The U. S. Preventive Services Task Force attached importance to the good internal consistency demonstrated by the HITS and WEB. They did not take into account that evidence based on internal structure does not reveal what questions are actually identifying (see section 1.3.6.3.); and that as the HITS is a bi-dimensional scale internal consistency reliability had not been used appropriately. The WEB uses a version of the safety question,[208] versions of which have also been used in the PVS[210, 211] and recommended by a systematic review to identify IPV.[231] Evidence from 1997 shows that the PVS safety questions have a PPV of 51%.[210] Feldhaus and colleagues interpreted their results as confirming that the PVS can detect a large number of women who have a history of IPV. Now in the context of the HARK questions achieving PPVs of over 79%, the PVS's PPV of 51% (whilst changing pre to post test probability by only 27%) appears to be too low.

A third systematic review from 2009 also concluded that the HITS had the best reliability, predictive power and concurrent validity with a suitable cut-off score.[7] Again there was no mention that as the HITS is not uni-dimensional that it was inappropriate to measure internal consistency reliability using Cronbach's alpha. This review also did not take into account the use of unacceptable reference standards by the three studies that investigated the HITS.[198, 215, 222] There was no recognition that in one study the very high sensitivity and specificity are likely to have been artificially inflated, having been derived from a known group comparison.[215] This systematic review did acknowledge that the HITS did not ask about sexual abuse and stated that this

could be overcome by using another tool to detect sexual IPV. I would argue that HITS' poor evidence of validity based on its content makes it inappropriate to use in clinical practice to identify IPV. This is especially so in the current context of tools that do achieve good evidence of validity based on content by including an enquiry about sexual IPV, i.e. the HARK questions,[200] the BRFSS questions[217] and the 8 question WAST.[211, 220]

The HARK study having been published in 2007 did not feature in the three systematic reviews examined above.[7, 50, 231] The most recent systematic review from 2009 did include the HARK, rating it as a good quality study, along with 13 other studies.[232] Overall this systematic review concluded that "No single IPV tool had well-established psychometric properties." It was noted that all the IPV tools needed additional reliability and validity testing. However like the three systematic reviews discussed above, existing numerical reliability and validity data were again taken at face value without unravelling their meaning. For example, for the HITS the same psychometric results, sensitivity and specificity were presented as in the systematic review above.[7] The erroneous use of internal consistency reliability and the impact of known group comparisons on results were not highlighted. The lack of consensus about appropriate reference standards was acknowledged in the discussion

I now reflect on the potential clinical importance of my research findings and how they could influence clinical practice. My systematic review did not find any questions that could be used in specific ethnic groups to identify IPV clinically. Unlike the systematic review, my secondary data analysis does provide evidence that the four HARK questions can identify IPV in self-classified UK census categories of south Asian, African-Caribbean and white groups. This is important as it means that on the basis of the existing evidence, the same questions can be used to identify IPV in individuals from different English speaking ethnic groups in primary care, in east London. Knowing that HARK has the same cut off in all three groups with high diagnostic indices is of note clinically as it allows HARK to be used in the same way in these groups. Clinicians and researchers of IPV can be reassured that there is no evidence from my systematic review and secondary

data analysis that the validity of questions varies significantly between different English speaking ethnic groups. A cluster randomised trial testing an educational intervention to improve the health care response to domestic violence used the four HARK questions as an electronic prompt to ask about IPV in response to given clinical presentations, for example depression and pelvic pain.[234] In this study (IRIS), the HARK questions have been found to serve as a reminder to clinicians to ask about IPV in multiethnic patient populations in Bristol and Hackney (personal communication). These clinicians and researchers can now be more confident that HARK has some validity in a variety of different ethnic groups.

Though the change from pre- to post-test probability of IPV produced by the kick question was greater than when using the four combined HARK questions, the kick question's sensitivity and NPV was lower. This would make it less likely for the four HARK questions to be replaced with just the kick question in the real world, despite Sackett's assertion about the supremacy of predictive values over sensitivity and specificity.[61] However in time limited scenarios, for example in emergency departments, it may be reasonable to use the kick question, as opposed to all four HARK questions. Conversely using the four HARK questions, starting with the humiliation question, may be interpreted as a gentler introduction to the difficult subject of IPV. In African-Caribbean women the first question when exploring the possibility of IPV being present may prove to be the afraid question which could then be followed by the kick question.

I have judged my research findings in the context of previous research and considered the potential clinical implications of my results. I now evaluate my appraisal of the quality of the methodology and the use of the ethnicity data in the 20 primary studies included in the systematic review and my secondary data analysis.

4.5. Quality Appraisal

In this section, I consider the role of QUADAS in evaluating the methodology of the 20 individual studies in my systematic review. The outstanding methodological issues related to the heterogeneity and hierarchy of the range of methods seen in these 20 systematic review papers as a whole are then also considered. Following this, I discuss my use of a simple checklist to appraise the use of ethnicity data in these 20 papers. Subsequently I evaluate the methodology of my secondary data analysis, focussing on the power of this analysis which is related to the use of ethnicity data and the use of a good reference standard.

4.5.1. Evaluating quality appraisal of methodology of the systematic review studies by QUADAS

The role of QUADAS in evaluating the methodology of the individual studies in my systematic review is considered below by highlighting QUADAS items which differentiate between studies and reflecting on QUADAS items which are not self explanatory. Further detail is given about these items to help contextualise them. See table 5 (page 127) for the QUADAS quality items in relation to each study.

The first QUADAS item, whether the spectrum of patients is representative, identified that three of the studies were known group comparisons with extreme groups[215, 219, 221] whilst a fourth involved high risk post partum women.[225] The three known group comparison studies were fundamentally different from those with more representative populations. This not only affected the degree of correlation and internal consistency reliability but also the diagnostic accuracy indices generated (see section 3.2.4.2.). This makes these studies incomparable with studies that do actually use a representative

spectrum of patients. Additionally the external validity of these studies is limited. The results are less generalisable to a general clinical population, including those women who have experienced varying degrees of IPV and have not required the services of a refuge shelter.

The second item, whether inclusion criteria have been stated, identified three studies which did not state inclusion criteria.[212, 213, 224] One of these did state exclusion criteria but the selection of participants was still not fully described.[213] QUADAS does not evaluate whether inclusion criteria are either appropriate or justified. For example, in the Zink study an inclusion criterion was to have been with a steady partner for at least one year.[216] This probably accounted for 81% of the study population being married. This may have decreased the prevalence of IPV in that study population (11%) which would have affected the PPVs and NPVs but not the low sensitivity (45%) of the index questions.

The third item, whether an acceptable reference standard was used, reveals that there were at least seven studies that definitely did not use an acceptable reference standard[208, 210, 215, 217, 222, 223, 225] with five studies in which it was unclear whether an acceptable reference standard was used.[198, 218-221] Hence there were eight studies that did use only an acceptable reference standard.[200, 209, 211-214, 216, 224] An acceptable reference standard is one which is likely to correctly classify the target condition (i.e. IPV). An apparent reference standard for IPV once modified requires further evidence that it remains an acceptable reference standard. Multiple modified comparators in the systematic review papers had not been previously psychometrically tested. For example, in Peralta's study[214] the modified reference standard (a six question version of the Conflict Tactic Scale) made it impossible to interpret the poor diagnostic accuracy indices generated. Instead this study should either have used an established reference standard or considered the use of other research methods that were not reliant on having a reference standard (for example convergent validity). However, it is clear that within the research area of IPV there is no universal consensus on a reference standard to measure IPV.

The modified comparators included those that were translated into different languages, for example Tiwari and colleagues[212] used the 39 question revised Chinese CTS (CTS2) whilst Reichenheim and colleague[213] used the 12 question physical violence scale of the modified revised Portuguese CTS2 (12 item mCTS2). In these instances I was reliant on the authors' comments having been unable to access the supporting references. Some of these studies also included index questions that had been translated into different languages. For example McFarlane and colleagues[224] offered the index questions in English and Spanish. No details were given on how the content validation of the Spanish AAS was achieved. It was difficult to say whether there was enough detail to replicate the foreign language versions of index and reference standard questions.

The tenth and eleventh items were concerned with review bias which refers to whether the index test results were interpreted without knowledge of the reference standard results (one QUADAS item) and whether the reference standard results were interpreted without knowledge of the index test results (further QUADAS item). This blind analysis of the index tool and the reference standard respectively is equivalent to "blinding" in traditional intervention studies. This is in an attempt to avoid review bias which could boost diagnostic accuracy measures. Table 5 shows that for all the studies bar one it was not known whether review bias was avoided. In the one study where analysis was known to be blinded, this information was not actually reported in the published paper but was acquired from the authors and confirmed from the original study protocol. However in all 20 studies, adding up individual's scores for sets of questions is an objective exercise which involves no subjectivity and is not vulnerable to review bias. Therefore it would have been safe to have omitted the two items concerned with review bias from QUADAS for this systematic review. For further details, see section 4.6.

Having covered the role of QUADAS in evaluating the methodology of the 20 individual studies in my systematic review, I now consider some outstanding methodological issues.

4.5.2. Outstanding methodological issues related to the systematic review

I now consider the methodological issues related to my systematic review which were not covered by the QUADAS items. This firstly includes the heterogeneity of methods followed by a proposal for the hierarchy seen in the methods present in the 20 papers identified in the systematic review.

4.5.2.1. Heterogeneity of methods

The Results chapter showed that the primary studies in the systematic review reported a variety of methods, trying to find evidence of validity for questions to identify IPV (i.e. were questions thought to identify IPV, identifying IPV?). These methods included diagnostic accuracy as well as traditional validation methods (criterion correlation, convergent validity, known group comparisons and internal consistency reliability). The heterogeneity of methods made it difficult to compare studies.

The inclusion criteria for my systematic review reduced the heterogeneity of study methods partly by stipulating that the comparator tool was either a standard reference criterion or other test intended to measure a construct similar or related to IPV (see section 2.2.2.). This in effect excluded studies that may have compared their index questions to external variables as opposed to a comparator tool. For example, in Coker's study[218] the index questions were also compared to self reported poor mental health and the number of physician visits in the last year. This type of association study between index scores and external variables (see section 1.3.6.2.2.1.) is reasonable especially when comparator tools are contested. Coker's study was included in my systematic review as the method also included comparison of the index questions to a comparator tool. There were other studies that were excluded as there was no comparison of the index questions to a comparator tool. For example, a known group comparison in which

there was no comparison of the translated Spanish WAST to a comparator tool was excluded from my systematic review.[235] On the one hand this was an asset, as decreasing heterogeneity of methods makes it easier to compare studies. Conversely this could be construed as a limitation excluding potentially important studies which may have produced evidence of validity for questions in a difficult arena in which there is no consensus about the ideal reference standard (see third QUADAS item, section 4.5.1.).

My computation of diagnostic accuracy in two studies[217, 218] was potentially a useful strategy to aid comparisons between studies. A previous systematic review also used this strategy.[7] The diagnostic accuracy data generated were easier to interpret than correlation data (i.e. coefficients, relative risks or overlapping numbers and percentages[1]) with regards to deciding whether a set of questions should be used in a clinical context based on their accuracy, as well as the questions then being easier to interpret when used in clinical contexts. The limitation of diagnostic accuracy data is the need for an efficient and appropriate reference standard and adequate categorical data (see section 4.5.2.2.). The WEB and modified ISA-P (used in the Bonomi and Coker studies respectively) are not dissimilar to the comparators treated as standard criterions from the onset in some studies. For example, a modified version of the ISA-P was used as a standard criterion in the study conducted by Chen and colleagues.[198] However it was incorrect to force comparator tools (WEB and modified ISA-P) to take on the function of reference standards when their utility as reference standards has not been confirmed. The Coker study data suggested that WEB's positive predictive value for identifying IPV was only 52%. This in turn made it unfeasible to interpret the diagnostic accuracy data that had been generated for the BRFSS. As well as being an inefficient reference standard, the WEB was probably also an inappropriate comparator tool. The authors pointed out that:

"…both the BRFSS and WEB identified some women as abused that would have been missed by the other instrument..."

[1] The numbers and percentages of women who scored positive on two tools, i.e. overlapped, were provided as a result in the Bonomi and Coker study.

Hence the BRFFS and the WEB may have been measuring different things, probably different aspects of IPV, making it inappropriate to use the WEB as a comparator tool to assess the performance of the BRFSS.

For my systematic review, I had planned to pool numbers by combining the results from studies about the same index questions in specific ethnic groups. This was not possible due to the heterogeneity of the comparator tools used (the majority of which were modified in different ways), the heterogeneity of the index questions used (for example, even when studies said that they were using the AAS in the Abstract, this invariably turned out to be different versions of the AAS questions) and the heterogeneity of the methods used, as described above.

4.5.2.2. Hierarchy of methods

Diagnostic accuracy studies do appear to be superior to criterion correlation studies within the subcategory of criterion performance studies. This is supported by my comparison of studies using both diagnostic accuracy and criterion correlation, in the Results (see section 3.2.2.8.). However criterion correlation studies have an important role if data cannot be used categorically, (see section 1.3.4.). A diagnostic accuracy method cannot then be used to assess the validity of questions. The HARK study data generated an optimal cut-off point in its entire study population as well as each of the three ethnic groups. This is evidence for the categorical rather than dimensional structure of the data.

I have discussed the heterogeneity and hierarchy of the range of methods seen in these 20 systematic review papers. I now consider my use of a simple checklist to appraise the use of ethnicity data in these 20 papers.

4.5.3. Evaluating quality appraisal of the use of ethnicity data in the systematic review studies by a five criterion checklist (DECSS)

The role of my five criterion checklist (DECSS) in evaluating the use of ethnicity data in the individual studies in my systematic review is considered below. See table 6 (Results, page 131) for these criteria in relation to each study.

I derived the individual criteria from published guidance on the use of ethnicity in health research. The role of this checklist was to appraise the use of ethnicity data in the primary studies included in the systematic review. These criteria essentially examine the ethnicity terms used in papers, how they are used and check to see whether confounding of ethnicity by socioeconomic status has been considered. See below.

Five criterion checklist (DECSS) for quality appraisal of the use of ethnicity data

1. **D**: Is ethnicity **described**?
2. **E**: What are the terms used to describe **ethnicity**?
3. **C:** Is the **classification** system using ethnicity justified?
4. **S**: Is ethnicity **self**-assigned?
5. **S**: If the study analyses differences in ethnic groups are **socio-economic** factors considered or controlled for?

Table 6 shows that out of the total of 20 studies in my systematic review, 18 described the ethnicity of the study population in some fashion. This table displays the complexity of ethnicity which can be defined in many ways (for example, using language, geographical origin, national census categories – see section 1.4.1.). This can make it difficult to have a straightforward answer to what at first seems like a simple question:

does a study describe ethnicity? This confusion is seen in some of the primary studies in my systematic review. For example, Chen and colleagues[198] conflate culture with language and socio-economic status whilst describing "Race / ethnicity."

It was striking that despite the terms being used ("ethnicity," "race / ethnicity," "race" etc.) the same group names were generally being used (white, black, Hispanic etc.). See table 2 (box headed "ethnicity") on page 115 and table 6 on page 131. Four studies used the phrase "race / ethnicity,"[198, 216, 217, 222] whilst three studies used the words "ethnicity" and "race" interchangeably.[208, 210, 224] Two studies used the word "race" but used the same group names as in studies that explicitly were purported to be studying ethnicity (i.e. African American, Hispanic, Asian, white).[214, 218] The majority of these classification systems appear to be based on skin colour with some containing additional information on country of origin and language. This supports the observation in my background (see sections 1.4.1.1. and 1.4.1.2.) that in published papers, there is little difference in how "ethnicity," "race" or "race / ethnicity" are used.

In my background I also suggested that the term "race" is outmoded as words such as Africans, Asians and Caucasians do not relate to any distinct genetic differences in humans.[140] The papers included in my systematic review show that these racial categories persist within biomedicine[140] as they do in every day life. For example, Ernst at al's relatively recent study from 2004[214] still used the misleading word "Caucasian." Bhopal points out that this is widely used as a synonym for "white" though it actually means originating in the Caucasus region, referring to Indo-Europeans.[134]

None of the studies justified the classification systems that they used. One study used national census categories to check the representativeness of its study population,[200] i.e. there was direct use made of the classification system.

In seven studies ethnicity was self-assigned.[200, 208, 210, 214, 217, 222, 224] In this last study, it was unclear what happened to participants who attended for an interview and whether their ethnicity was still self-assigned. In the remaining studies it was not

clear whether ethnicity was self-assigned or assigned by others. For ethnic classifications to be valid they should preferably be self assigned, (see section 1.4.1.).

Out of the 12 studies whose study populations contained different groups, four studies analysed the differences between groups.[198, 208, 216, 224] Two of these studies compared the evidence for validity of questions trying to identify IPV in different ethnic groups with consideration of socio-economic status.[198, 216] One of these two studies argued that it had found a difference between groups but did not formally check for confounding by socioeconomic status.[198]

The majority of the studies whose study populations contained different groups did not analyse the differences between groups. There was also no further processing of their ethnicity data apart from in one study out of the eight.[200] Collecting ethnicity data without analysing it further adds to the description of the study population, along with the participants' age, sex and socioeconomic status. This helps the reader judge how pertinent any particular research finding is to their own clinical population. However the downside of the almost ubiquitous presence of unprocessed ethnicity data in IPV identification studies is that it creates an impression that an ethnic difference does exist in IPV identification. This is in contrast to the actual findings of my research that there is nothing to suggest a difference in IPV identification between ethnic groups. If researchers believe that ethnicity impacts on IPV identification, their studies should be organised so as to measure its impact. If researchers believe that ethnicity does not impact on IPV identification but their studies contain diverse populations to increase the external validity of their findings then the impact of ethnicity does not need to be measured but the ethnic profile of the sample should be characterised to assess the extent to which it is representative of the local general population. Epstein expanding on his Inclusion / Difference paradigm affirms that when including a diverse study population one may want to see if it is representative of the actual population but this should not mean that one is looking for an ethnic difference in the study variable being studied unless one has a pre-existing premise supporting an ethnic difference.[140]

In my background, my premise was that cultural differences in attitudes towards IPV could affect disclosure in different ethnic groups which in turn could also affect how accurately some questions identify IPV in different ethnic groups in a health setting (see section 1.4.1.3.). Therefore from my initial perspective, considering that the majority of studies contained both ethnicity and socio-economic data, the fact that only four studies analysed any differences in ethnic groups is a missed opportunity for investigators who could have compared the evidence for validity of questions trying to identify IPV in different ethnic groups. However there are resource implications for a study to be sufficiently powered to look at ethnic differences. This may explain why most studies have not analysed the ethnic differences in IPV identification.

In conclusion, I consider that my list of five set criteria (DECSS) has allowed the task of appraising the use of ethnicity data in the systematic review studies to be achieved in a standardised manner. This short checklist is easy to apply to studies unlike the existing lengthy published guidance.[137, 148, 165, 166, Malley-Morrison, 2007 #440] Published guidance about the use of ethnicity appears to be mostly ignored in the topic area of IPV identification. It has also been overlooked further afield.[236]

4.5.4. Evaluating the secondary data analysis – methodology and use of ethnicity data

I now consider the methodological issues and the use of ethnicity data in my secondary data analysis. Firstly, the major methodological strength of my secondary analysis was that I calculated confidence intervals for all my results which provided a measure of the precision of my results, unlike the majority of the studies included in the systematic review. The confidence intervals suggested that my secondary analysis was underpowered to confidently confirm that there were no statistically significant differences in the validity of the HARK questions at identifying IPV in different ethnic groups.

In the original HARK study, a power calculation established the size of the study population required.[200] A power calculation was not possible for my secondary analysis as the size of the study population was already fixed with predetermined numbers of participants in each ethnic group. The systematic review showed that a robust investigation into the role of ethnicity on IPV identification had not previously been conducted. This indicated that my secondary analysis was of value, as it would be able to inform the design of a future definitive study into the ethnic differences in IPV identification.

This future study would need a total sample size of 2,142 (i.e. 238 in each group) if there were nine different ethnic groups. These nine groups would be more specifically defined than the current three groups. This study could then potentially show that clinically important differences in the validity of the HARK questions at identifying IPV is unlikely – assuming that clinically important differences are in the order of 20%, with study power being 80%. This sample size calculation is based on the assumption that the change from pre-test to post-test probability of IPV in all groups is 50% which is consistent with the current data.

The second methodological strength of my secondary analysis was that I used the Composite Abuse Scale (CAS) as my reference standard. Based on published evidence this appears to be a good reference standard for measuring IPV.[101, 115] In my original HARK paper, I state that

"The CAS has an internal reliability (Cronbach's alpha) of .90 or more for each sub-scale, and all item-total score correlations of .6 or above. It has also been validated with a large (1,836) sample of patients in general practice settings within primary care. It is based on a concept of IPV that includes coercion, not simply violent acts arising out of conflict. It is recommended as an IPV research assessment tool by the National Centre for Injury Prevention and Control as it has demonstrated reliability and validity for measuring the self-reported incidence and prevalence of IPV. It has evidence of content, construct, criterion and factorial validity."[200]

The CAS represents self-reported verified measurable violence unlike existing tools.[50] I believe that on the basis of this evidence the CAS deserves a consensus recognising it as the best current reference standard with which to measure IPV. This consensus currently does not exist (see third QUADAS criterion, section 5.5.1.). Evaluating the HARK questions against the CAS allows HARK's accuracy for identifying IPV among women in the general population to be demonstrated. CAS's most important limitation is that it does not have a record of being tested in the ethnic groups in which I was trying to validate the HARK questions. However neither did any of the other acceptable reference standards.

In my secondary analysis I amalgamated data into larger groups (African-Caribbean, south Asian and white) in an attempt to increase power to make comparisons between groups. I felt that this was justified as the participants within each aggregated group may share cultural beliefs affecting how they may respond to questions asking about IPV, perhaps related to their historical connections with distinct geographical regions, i.e. Africa, the Indian subcontinent and Europe respectively. The advantage of using these aggregated groups was that this increased the power of my analysis to make comparisons between groups without increasing the resources that I needed. The limitation of this approach was the accompanying assumption that the cultural beliefs shared by women within each group (for example a first generation Nigerian immigrant with a second generation Jamaican woman) had more in common than the cultural beliefs shared by a woman from one group with a woman from one of the other two groups (for example, the Nigerian woman with a white English woman). This type of assumption is supported by qualitative research that suggests racism may affect abused individuals' responses in surveys and their experiences of abuse.[237] Racism has also been intimated to be used by men to gain their partner's forgiveness[238] with the role of black-led churches at times of personal distress recognised.[238, 239] However this should not distract from the differences within these groups – for example, African communities have often recently immigrated to London compared to more settled Caribbean communities.

Studies of ethnic differences always have to balance studying specific well defined ethnic groups versus resource constraints requiring the study of larger less well defined groups. These larger groups often end up describing participants' skin colour and / or continent of origin rather than their ethnicity. Hence the groups in the secondary analysis also appear to be based on skin colour despite the HARK study being based on national census categories of ethnicity and the background to this thesis making a case for IPV identification being affected by culture as opposed to skin colour.

The wide overlapping confidence intervals generated for the diagnostic indices for each aggregated ethnic group (see tables 7, 8 and 9) means that the size of each group was still too small and / or that the differences between the groups were of an order that were too slight to be detected by my data analysis. The results may be consistent with there being no real ethnic differences in the validity of questions to identify IPV. This would be confirmed by having a sufficient number of participants in each ethnic group and the generation of overlapping narrow confidence intervals for the diagnostic indices.

I have considered how I appraised the quality of the methodology and of the use of ethnicity data in the 20 primary studies included in the systematic review and my secondary data analysis. I now turn back to QUADAS, an evidence based quality assessment tool. In the next section, I consider QUADAS's limitations which were highlighted when I used it to appraise my systematic review. I then compare QUADAS to the Standards for Reporting of Diagnostic Accuracy (STARD) statement.

4.6. Limitations of QUADAS

Quality appraisal of the primary studies with QUADAS highlighted important methodological weaknesses and biases (see section 4.5.1.). However some limitations of QUADAS were also isolated in that not all of the methodological problems in studies were identified. I now consider some of the limitations of QUADAS and then compare QUADAS to the *Sta*ndards for *R*eporting of *D*iagnostic Accuracy (STARD) statement. STARD arose more directly from the Consolidated Standards of Reporting Trials (CONSORT) initiative.[240] It was developed by a group of scientists and editors in 2003.[241-243]

QUADAS has no item for sample size or whether a power calculation has been used to estimate the sample size required. Hence a number of studies with extremely small sample sizes[219, 221] were not down graded and studies with large sample sizes were not upgraded.[211, 217] QUADAS also does not have a criterion to evaluate whether statistical tests are used correctly. For example only categorical data permits diagnostic accuracy indices to be calculated. There was also no criterion addressing the importance of using confidence intervals to assess the precision of results. I think that an improved QUADAS should contain items that:

1. judge adequacy of sample size in relation to study aims
2. evaluate whether statistical tests are used correctly
3. recognise the use of confidence intervals

The initial list of 28 possible items for inclusion in QUADAS included[202]:

- Was an appropriate sample size calculation performed and were sufficient patients included in the study?
- Were appropriate results presented (sensitivity, specificity, likelihood ratios, diagnostic odds ratios and predictive values) and were these calculated appropriately?

- Was a measure of precision of the results presented (confidence intervals, standard errors)?

These 28 items were rated using a consensus Delphi procedure which finally resulted in 14 items.[202] The process of the Delphi procedure was described in detail. The precise reasons why particular items were excluded were not stated. A summary of evidence from a review of methodological literature on diagnostic test assessment and a review of the tools used to assess the quality of diagnostic tests in relation to each of these items was given to the diagnostic experts who comprised the panel in the Delphi procedure. This was to assist the panel members in their task of rating each item for inclusion in QUADAS, the quality assessment tool.

For both the excluded items on sample size calculation and results presented (see above), the review of methodological literature on diagnostic test assessment showed that no studies were available providing evidence of either empirical or theoretical evidence of bias or an absence of bias and variation. The review of the tools used to assess the quality of diagnostic tests showed that between 25 to 49% of these tools covered both appropriate sample size calculation and the results presented.

For the excluded item on precision of results, the review of methodological literature on diagnostic test assessment showed that there were no studies providing evidence of either empirical or theoretical evidence of bias, or evidence of an absence of bias and variation. 0 to 24% of tools used to assess the quality of diagnostic tests covered this item on the precision of results.

The Delphi procedure received completed questionnaires from eight of the eleven diagnostic experts initially invited. Five panel members endorsed the Delphi procedure. One did not. He stated: “I fundamentally believe that it is not possible to develop a reliable discriminatory diagnostic assessment tool that will apply to all, or even the majority of diagnostic test studies.” The two remaining panel members were unclear about whether to endorse the Delphi procedure.[202]

In common with the previous research done into QUADAS, I also found that a combined quality score should not be used.[205] The range seen in the number of items fulfilled by each study included in the systematic review varied from eight to 12 when the two items concerned with review bias were removed (see section 4.5.1.). This limited range did not demonstrate the actual wider variety uncovered in study quality. For example, the McFarlane study fulfilled ten items which would initially appear to indicate reasonable quality. I think that the un-interpretable results (no numerical data presented) are more important with respect to study quality than the other ten items. This supports the view that a combined quality score should not be used due to unresolved weighting issues.[205]

However the systematic review also showed examples where the range seen in the number of QUADAS items fulfilled did seem to indicate something tangible. For example examining the two studies which reported differences in validity evidence between ethnic groups,[198, 216] the first fulfilled nine items whilst the second fulfilled twelve items. This difference seems to represent a true difference in the overall methodological quality of these two studies. Though the Zink study fulfilled the maximum number of QUADAS quality items it still did not support the use of its five non-graphic questions for identifying IPV in either white or African American populations. Indeed having fulfilled the maximum number of quality items, adds weight to the fact that these questions are not valid in these populations.

Having considered the limitations of QUADAS, I now compare QUADAS to the STARD statement.

4.6.1. Comparison of QUADAS to STARD

The objective of QUADAS is to assess studies of diagnostic accuracy included in systematic reviews whilst the objectives of STARD are to improve the quality (i.e. accuracy and completeness) of the reporting of diagnostic accuracy studies, allowing the

bias potential and generalisability of a study to be assessed.[241] Unlike QUADAS, STARD includes criteria to

- describe the methods for calculating or comparing measures of diagnostic accuracy, and the statistical methods used to quantify uncertainty (for example, 95% confidence intervals) - which would be affected by the sample size.
- report estimates of diagnostic accuracy and measures of statistical uncertainty (for example, 95% confidence intervals).

These issues have not been addressed by QUADAS. Additionally STARD touches upon test reliability with items on reproducibility[2]

- Describe methods for calculating test reproducibility, if done.
- Report estimates of test reproducibility, if done.

QUADAS contained no items to look at these characteristics in the index tool although incongruously further planned work to assess QUADAS includes checking its consistency and reliability.[202]

For STARD, as for QUADAS, the literature was searched to find 75 potential items but unlike QUADAS a two day consensus meeting was used to reduce this to a 25 item checklist and a flow diagram, using evidence whenever it was available. The STARD consensus meeting was attended by 39 specialists (STARD steering committee, 9 members and STARD group, 30 members), resulting in every-one signing up to the final STARD statement. Overall I think that STARD appears to have a number of advantages over QUADAS including greater professional endorsement.

However neither QUADAS nor the STARD are designed to appraise studies that predominantly use validation paradigm methods. Hence the fact that correlation coefficients and Cronbach's alphas were not used appropriately in different studies would not have been identified by either. Hardly any of the papers using correlation coefficients presented their scatter plots. The meaning of any particular correlation coefficient is

[2] Reliability refers to the reproducibility of a measurement.

assisted by presenting the associated scatter plot displaying the bivariate distribution between the test score and other measure – a point not captured by QUADAS or STARD. Additionally neither QUADAS nor STARD have items assessing the validity evidence based on response processes or questions' content. In traditional diagnostic accuracy studies concerned with medical tests this may be irrelevant but when assessing questions that form part of a history these characteristics become central, especially if questions have been translated for evaluation in different ethnic groups.

QUADAS is constrained because it is a single tool trying to apply to all diagnostic accuracy studies. It has been described as the generic part of a more extensive tool which would include topic specific items, for example for questionnaire scales.[202] Indeed it may be that a separate new appraisal tool is required to assess the quality of studies in systematic reviews that apply the concept of clinical test validation. This is when researchers take the view that no acceptable reference standard exists,[3] and validation paradigm methods are used. Topic specific items may then include checking that correlation coefficients and Cronbach's alpha are used appropriately. This assessment requires a clear description of the theoretical basis of a construct that a questionnaire is trying to tap. Items assessing the evidence of validity based on response processes or questions' content would also be useful, especially if examining questions in different languages whilst looking at ethnic differences.

I have considered QUADAS's limitations whilst comparing it to the STARD statement and discussed how neither QUADAS nor STARD have been designed to appraise studies that predominantly use validation paradigm methods. I now discuss the strengths and limitations of my thesis.

4.7. Strengths and limitations of my thesis

I now discuss the strengths and limitations of my thesis as a whole, as opposed to those covered above which mainly relate separately to the systematic review and secondary analysis.

4.7.1. Strengths

The strengths of my thesis include that it directly builds on the work previously carried out in identifying IPV, that it uses my comprehensive categorisation of evidence of validity which conveys a lucid understanding of validity, my attention to detail with a focus on first principles, my highlighting of the ability of index questions to alter the probability of IPV, my use of a checklist (DECCS) to appraise the use of ethnicity data in studies and my recognition that all individuals have an ethnic identity. I expand on each of these strengths below.

The foremost strength of my thesis is that it directly builds on the work previously carried out in identifying IPV whilst also contributing to future research. The original HARK study was planned after reviewing the index questions in the literature being used to identify IPV. Thus the HARK questions arose from the AAS questions.[200] The original HARK paper includes a full description of how the HARK questions were adapted from the AAS questions with discussion about the AAS questions' strengths and weaknesses. The HARK questions attempted to build on the strengths of the AAS whilst eliminating its weaknesses. Following this non-systematic review of index questions, my work evolved into a systematic review of index questions to identify IPV in specific ethnic groups with my accompanying secondary data analysis. My thesis will now be able to contribute to a potentially authoritative study which precisely defines questions that are valid in some specific ethnic groups to identify IPV as well as being able to confirm that

there are not any significant differences in the validity of these questions between ethnic groups (see section 4.5.4. and 5.3.).

My comprehensive categorisation of evidence of validity (see section 1.3.6.) is also a major strength of my thesis as it allowed me to thoroughly review all the evidence of validity presented in the papers included in my systematic review as well as my secondary analysis. This framework also accommodated evidence of validity from papers that did not fulfil the inclusion criteria of my systematic review (for example, Fogarty and Brown's study[235]). I devised this framework by integrating the evidence derived from diagnostic accuracy studies and more traditional validation methods including validity evidence based on response processes and test content. None of the previous four systematic reviews had looked methodically at these latter two attributes of index questions. I think that validity evidence based on response processes and test content are fundamental factors to consider when evaluating index questions. Most recently Streiner and Norman noted that the Standards for Reporting of Diagnostic Accuracy (STARD) statement and the Standards for educational and psychological testing were useful guidelines for reporting test results.[103] I believe that my integration of diagnostic accuracy methods and traditional validation methods goes an important step further by considering together in one framework how validity can be measured and how different sources of validity potentially relate to one another.

My clarity regarding the concept of validity is also a thesis strength. This is reflected by the use of up to date terminology i.e. that all validity is construct validity as opposed to the traditional three Cs of validity (criterion, construct and content validity). Following on from this is that validity (an outcome) is very different from validation (a process). Linked to this is that validation is an ongoing process which changes the degree of confidence that one draws about the inferences made about participants according to the scores they have obtained on answering index questions. This in turn highlights the importance of using confidence intervals which convey the precision of the result. Confidence intervals featured prominently in the original HARK study and my secondary data analysis unlike in the majority of the primary studies in my systematic review.

A further strength of my work has been my attention to detail with a readiness to return to first principles for each subject covered (for example, measuring validity and defining ethnicity). This was also seen in relation to the statistical methods used in studies. Hence I consistently considered how various statistics were used rather than taking numerical results at face value. This has led to my questioning of results that other systematic reviewers have accepted.

The importance that I have attached to the ability of index questions to change the pre-test probability of IPV to a different post-test probability of IPV is also a strength. I stated that the most clinically useful index questions are those that are found to produce the largest difference between the specific study prevalence and the PPV of the index questions in that study. The rationale behind this has been explained in the Background (see section 1.3.6.2.1.1.). Therefore the sensitivity of index questions for identifying IPV is not paramount as proposed by some[232] nor are just the predictive values as suggested by others.[61] I think that using the difference between pre- to post-test probabilities gives the most useful information about the functional ability of index questions to identify IPV. It also underscores that it is not questions that are valid. Instead validity applies to the application of questions in a particular study population, with its own unique IPV prevalence and not to the questions them-selves.[89]

A further strength is my compilation of five set criteria (DECSS) to appraise the use of ethnicity data in studies. This checklist (DECSS) has potential use for researchers planning studies into ethnicity, for peer reviewers evaluating studies about ethnicity and for publishers deciding whether to publish studies on ethnicity. This in turn may serve to increase the quality of studies by biomedical researchers examining ethnicity and ethnic differences or at least to improve the quality of the reporting of these studies.

Lastly, a fundamental strength of my research is the underlying concept that all individuals have an ethnic identity. Empirically this translated into my systematic review having no inclusion or exclusion criteria that were concerned with ethnicity. Consequently some of the included studies examined ethnic majorities (for example,

Bonomi et al's study of white Americans[217]) as opposed to studies which only examined ethnic minorities. The Oxford Dictionary defines "ethnic minority" as a usually identifiable group differentiated from the main population of a community by racial origin or cultural background.[90] Historically many majority community researchers have studied only ethnic minorities which detracts not only from the ethnicity of other white minority groups (for example, the Irish in the UK) but also the ethnicity of the majority group. Most importantly presuming that all individuals have an ethnic identity offers a straightforward strategy for avoiding stigmatising and the "we / they" dichotomy between researchers (who are often from majority communities) and ethnic minorities. Stigmatising and the "we / they" dichotomy has been described as a major challenge when writing about ethnicity.[137] Other strategies to avoid the "we / they" dichotomy such as ensuring more researchers from minority communities (such as myself) are probably harder to achieve especially in the short term. Not using ethnicity as either an inclusion or exclusion criterion also resulted in more papers being included in the systematic review and a wider variety of validation methods being appraised which improved the robustness of inferences made. This was an additional strength.

4.7.2. Limitations

The limitations of my research are related to a lack of financial resources which resulted in not being able to use more than one reviewer of data, not being able to use an expert panel to assess validity evidence based on test content and a restricted examination of language and its relationship to ethnicity. I now expand on each of these limitations below.

Due to limited resources I was unable to use two reviewers to read eligible papers, record data, independently assess study eligibility and study quality for my systematic review. Consequently a third reviewer was also not required as judicator which is normally needed when the first two reviewers do not agree. Linked to this was also not having an expert panel to assess validation from the content of the index questions. Instead all of

these roles were conducted by me. It could be argued that I had conflicting interests as I was also the lead author of one of the papers included in the systematic review, especially as this paper was then also found to perform well with regards to QUADAS and other criteria (for example, validation according to its content). However due to my awareness of this limitation from the onset, I used objective criteria which were less vulnerable to biased interpretation by me. I would also consider that I acquired expertise in IPV identification through my lead role in developing the HARK questions and careful consideration of their content. This involved me judging the content of other index questions in order to ensure that the HARK questions built on the strengths of existing questions whilst avoiding their flaws. I think that this has increased my ability to judge validation for identifying IPV from the content of index questions.

A further limitation of my research was that language failed to be captured. The majority of the systematic review studies and my secondary analysis were conducted in English speaking ethnic groups in the developed world.[165, 200, 208-211, 214-220, 222] Women who did not speak English were mostly excluded by these studies. Yet language is an important component of ethnicity (see section 1.4.1.) and marker of acculturation which may be a factor responsible for ethnic differences in IPV identification. Globally language is also interpreted more consistently than ethnicity, whereas ethnicity at its root is a sociological concept interpreted differently around the world. For example, Brazil is an ethnically diverse society but studies from Brazil rarely describe or look at differences in ethnicity. This is related to multiple socio-political and historical factors.[244] Examining IPV identification in different language groups may expose key differences.

I have now discussed my results and their implications, considering them also within the context of my thesis' strengths and limitations. Next in my final chapter, I present my overall conclusions and consider future research required.

Chapter 5: Conclusions and Future Research

5.1. Overview

In this chapter, I will present the conclusions of my thesis. I will also reflect on the extent to which my research questions have been answered. This is followed by recommendations for future research required in this field. This includes measuring validity using different categories of evidence, improving future systematic reviews and methodological studies of quality appraisal tools, including the DECCS.

5.2. Conclusions

My thesis conclusions are first that the only questions shown to have some validity for identifying IPV in specific ethnic groups were the four HARK questions and second that no evidence suggested that the validity of questions used to identify IPV varied significantly between different ethnic groups.

This was based on my secondary data analysis that showed using a cut-off score of ≥ 1 the four HARK questions had high diagnostic indices for identifying IPV in self-classified UK census categories of African-Caribbean, south Asian, and white groups. In contrast the systematic review offered no evidence of questions that were valid for identifying IPV in specific ethnic groups, including white groups. Neither the systematic review nor the secondary data analysis provided any evidence that the validity of questions used to identify IPV varied significantly between different ethnic groups.

Thus my principal research question, (what is the evidence for the validity of questions trying to identify IPV in specific ethnic groups?) has been answered by my research. My second research question (does the evidence for the validity of questions trying to identify IPV vary between different ethnic groups) has been answered in that the current evidence for validity of questions trying to identify IPV has not been found to vary between different ethnic groups. However both of my two research questions have not been completely addressed as the evidence for the validity of questions trying to identify IPV in specific ethnic groups is limited.

5.3. Future research

Future studies should provide evidence for the validity of the HARK questions to identify IPV in specific ethnic groups, using other methods apart from diagnostic accuracy. My categorisation of evidence of validity (see section 1.3.6) demonstrates that a variety of different evidence derived from different methods would support the validity of the HARK questions. I now describe future studies that could be used to collect category A evidence (evidence for the validity of the HARK questions that is based on the consequences of testing), category B (evidence for the validity of the HARK questions that is based on relations to other variables), category D evidence (evidence for the validity of the HARK questions that is based on response processes) and category E evidence (evidence for the validity of the HARK questions that is based on test content). Category C evidence (evidence for the validity of the HARK questions that is based on internal structure) is also considered.

Category A evidence for index questions trying to identify IPV, including the HARK questions, is absent from the literature but would be the most useful evidence to generate from studies in the future. This recognises that despite the good diagnostic accuracy of the HARK questions at identifying IPV they are only of value if they improve outcomes for women experiencing IPV. This is consistent with the GRADE approach to grading the quality of evidence and strength of recommendations for diagnostic tests.[245]

An efficient way of generating category A evidence may be to employ studies which are already using the HARK questions and performing secondary analysis of their data. This includes the IRIS study which was based in the UK[234] as well as studies that are planned for two sites in the US and Germany, (personal communications). The IRIS study is the first European randomised controlled trial of an educational intervention to assess the health care response to domestic violence. The primary outcome is the referral of women to specialist domestic violence agencies. Intervention practices have had the HARK questions integrated into their electronic medical records. The Metro Alliance for

Healthy Families in the Minneapolis-St. Paul metropolitan area, in Dakota County, in the US plan to use the HARK questions whilst designing evidence based training for their health visitors. They aim to collect outcome measures in order to use evidence to implement, improve and expand their health visiting programme. Additionally a nurse researcher will be piloting the HARK questions in emergency rooms, in family practices, in an obstetrics and gynaecology practice and a student infirmary at a university in Upstate New York. If the pilot is satisfactory, the plan is to incorporate the HARK questions into the electronic health record of all patients in 25 practices. Researchers in Germany have translated the HARK questions into German in order to use it in research in Germany.

Secondary data analysis of existing trials may potentially lead to answers about the consequences of using the HARK questions, i.e. whether using the HARK questions leads to any improvement in the health outcomes for women who are identified as experiencing IPV in different ethnic groups. Careful consideration would need to be given regarding consent for the secondary analysis of data as well as the reliability of the recording of health outcomes and ethnicity on primary care computer systems.

Category B evidence could include further criterion performance studies in different settings. Assessing HARK's diagnostic accuracy outside of the UK in other developed regions (for example Australia and North America) as well as developing nations would be logical. Criterion correlation is an alternative method to diagnostic accuracy if the IPV construct is measured using a continuous as opposed to a categorical scale, (see section 1.3.4. and section 1.3.6.2.1.).

Further Category B evidence includes association studies examining the association between HARK scores and external variables. These could include the number in the preceding year of primary care consultations or referrals for specialist care or prescriptions issued. Association studies do not make an assumption that a reference standard exists for identifying IPV. It would be possible to look at the association between the use of HARK as an electronic prompt and external variables by using the

EMIS-WEB database of routinely collected information from primary care in north east London. This database allows one to identify the 877 women identified as disclosing IPV during the IRIS study in intervention and control practices.

Future studies should also examine the evidence of HARK's validity based on response processes and test content whilst using an independent panel of IPV experts who are not directly linked to the development of the HARK questions. Content validation could be achieved by generating a content validity index for each individual HARK question. This is the proportion of judges who rate the question on three or four when content relevance is represented by a four point scale. One symbolises totally irrelevant content and four extremely relevant content.[246] This process could simultaneously be carried out for other tools that I have classified as having satisfactory content validity (for example, the BRFSS), to allow a more objective quantitative comparison of index questions.

Future studies do not need to investigate the evidence of the HARK questions' validity based on internal structure. This is not indicated as the four HARK questions essentially embrace different subscales of IPV which is a multidimensional construct. Hence internal consistency need not be considered. Additionally neither inter observer reliability nor intra observer reliability need to be measured as the HARK score is not vulnerable to either inter or intra observer variation as adding up an individual's HARK score is an objective exercise. Test – retest reliability of the HARK questions may be useful but it would be difficult to assess the optimum time for re-administration of the HARK questions. Too short a time period may not be adequately testing the actual test – retest reliability whilst too long a period may result in changes in a woman's experience of IPV. Therefore changes in HARK scores may not actually reflect the HARK questions' true test-retest reliability.

It would be useful to update the current systematic review regularly, for example, every five years. A future systematic review could include a more heterogeneous range of methods examining validity including association studies. This would be indicated if the

literature of IPV identification continues to show a lack of consensus regarding a reference standard for identifying IPV.

Future systematic reviews of IPV identification would also be improved by having better tools to appraise the quality of studies. This would be achieved by organising methodological studies examining quality appraisal tools. For example, further research to develop an evidence based tool which is able to appraise the quality of studies using validation paradigm methods would be useful (see section 4.6.1.). As with QUADAS, this new appraisal tool should ideally be: "…systematically developed and evaluated…for its usability and validity.[202]" Therefore though I have suggested possible items for inclusion in this new tool, in section 4.6., these items should ideally be confirmed for inclusion by being identified using a mixture of empirical evidence (if any is found using a systematic review) and expert opinion. This process could be formalised using a Delphi procedure which focuses on consensus.

My five criterion checklist, DECSS that assessed the use of ethnicity data in IPV identification studies should ideally be scrutinised by established researchers of ethnicity and ethnic differences. It is unclear whether this is best achieved using a Delphi procedure or a face to face consensus meeting of recognized researchers. Whatever the precise process, this may ultimately result in a revised DECSS checklist which should have its function assessed in other subject areas apart from IPV identification.

A definitive study to answer my second research question (does the evidence for the validity of questions trying to identify IPV vary between different ethnic groups?) would need a sample size of 2,142 with 238 participants in each of nine ethnic groups to confirm that there are no statistically significant clinical differences in the validity of the HARK questions at identifying IPV in different ethnic groups. This would avoid using aggregated ethnic groups. If there were no real ethnic differences in the validity of questions to identify IPV, the diagnostic indices would have narrow overlapping confidence intervals.

It may be more productive for future studies to use language as a vehicle for investigating ethnicity. Technically these studies would be difficult to set up and resource – see the issues discussed in the background on translation of questions, section 1.3.6.6.1.. Costs may be minimised by the use of online data collection, as described by Sackett.[247] If differences between language groups were more marked than between groups based on national census categories, smaller numbers of study participants may be required which could reduce costs. Overall this could be cheaper than funding a study requiring 2,142 participants. In an increasingly competitive funding environment this is an important consideration.

A range of different types of evidence of validity in different ethnic groups would support the case for including the HARK questions' in the Quality and Outcomes Framework. This is an incentive programme for all general practice surgeries in England. However the strongest evidence for inclusion would be to show that by using the HARK questions to identify IPV, one was able to improve the health outcomes for women experiencing IPV, i.e. category A evidence, regardless of their ethnicity. Nonetheless health policies are not based solely on evidence.[248] I judge that IPV already deserves consideration for inclusion in the Quality and Outcomes Framework due to its importance as a public health problem affecting not only the long term health of women who experience it but also the children who witness it. This combined with these women being relatively isolated yet having repeated contact with primary care and wanting support from health professionals[72] supports inclusion.

My final conclusions are that my research is contributing to a gradual coming of age of evidence based diagnosis with the incorporation of validation methods and the study of ethnicity within the field of gender violence health related research. I have tried to show above how these processes in the future will lead to better research questions and answers for women experiencing IPV. This discourse should not divert individual health workers from always being open to spontaneous IPV disclosure, responding suitably to women disclosing IPV and providing a patient centred approach focussing on cultural

competence[152] whilst avoiding stereotypes of the ethnic groups from which individual women originate.

References

[1] Tjaden P, Thoennes N. Extent, nature and consequences of intimate partner violence: findings from the National Violence Against Women Survey. Retrieved December 22, 2008, from http://www.ncjrs.gov/txtfiles1/nij/181867.txt. Washington, DC: US Department of Justice; 2000.

[2] Heise LL. Violence against women: an integrated, ecological framework. Violence Against Women. 1998 Jun;4(3):262-90.

[3] Rutjes AW, Reitsma JB, Coomarasamy A, Khan KS, Bossuyt PM. Evaluation of diagnostic tests when there is no gold standard. A review of methods. Health Technol Assess. 2007 Dec;11(50):iii, ix-51.

[4] Watts C, Zimmerman C. Violence against women: global scope and magnitude. Lancet. 2002 Apr 6;359(9313):1232-7.

[5] World Health Organization. WHO multi-country study on women's health and domestic violence against women: summary report of initial results on prevalence, health outcomes and women's responses. Geneva; 2005.

[6] Garcia-Moreno C, Jansen HA, Ellsberg M, Heise L, Watts CH. Prevalence of intimate partner violence: findings from the WHO multi-country study on women's health and domestic violence. Lancet. 2006 Oct 7;368(9543):1260-9.

[7] Feder G, Ramsay J, Dunne D, Rose M, Arsene C, Norman R, et al. How far does screening women for domestic (partner) violence in different health-care settings meet criteria for a screening programme? Systematic reviews of nine UK National Screening Committee criteria. Health Technol Assess. 2009 Mar;13(16):iii-iv, xi-xiii, 1-113, 37-347.

[8] Finney A. Domestic violence, sexual assault and stalking: findings from the 2004/05 British Crime Survey; 2006 12/06.

[9] Richardson J, Coid J, Petruckevitch A, Chung WS, Moorey S, Feder G. Identifying domestic violence: cross sectional study in primary care. Bmj. 2002 Feb 2;324(7332):274.

[10] Hegarty K. What is intimate partner abuse and how common is it? In: Roberts G, Hegarty K, Feder G, eds. Intimate partner abuse and health professionals : new approaches to domestic violence. Edinburgh: Churchill Livingstone 2006:32-5.

[11] Howard LM, Trevillion K, Khalifeh H, Woodall A, Agnew-Davies R, Feder G. Domestic violence and severe psychiatric disorders: prevalence and interventions. Psychol Med. 2010 Jun;40(6):881-93.

[12] National Center for Injury Prevention and Control. Costs of Intimate Partner Violence Against Women in the United States. Atlanta, GA: Centers for Disease Control and Prevention; 2003.

[13] Richards P. Homicide statistics Research paper 99/56. 1999 [cited 2005 accessed 13 Mar]; Available from: http://www.parliament.uk/commons/lib/research/rp99/rp 99-056.pdf

[14] Ellsberg M, Jansen HA, Heise L, Watts CH, Garcia-Moreno C. Intimate partner violence and women's physical and mental health in the WHO multi-country study on

women's health and domestic violence: an observational study. Lancet. 2008 Apr 5;371(9619):1165-72.
[15] Campbell JC. Health consequences of intimate partner violence. Lancet. 2002 Apr 13;359(9314):1331-6.
[16] Vos T, Astbury J, Piers LS, Magnus A, Heenan M, Stanley L, et al. Measuring the impact of intimate partner violence on the health of women in Victoria, Australia. Bull World Health Organ. 2006 Sep;84(9):739-44.
[17] Coker AL, Davis KE, Arias I, Desai S, Sanderson M, Brandt HM, et al. Physical and mental health effects of intimate partner violence for men and women. Am J Prev Med. 2002 Nov;23(4):260-8.
[18] Yoshihama M, Horrocks J, Kamano S. The Role of Emotional Abuse in Intimate Partner Violence and Health Among Women in Yokohama, Japan. Am J Public Health. 2008 Aug 13.
[19] Ludermir AB, Lewis G, Valongueiro SA, de Araújo TV, Araya R. Violence against women by their intimate partner during pregnancy and postnatal depression: a prospective cohort study. Lancet. 2010 Sept 11;376(9744):851-2.
[20] Campbell R, Greeson MR, Bybee D, Raja S. The co-occurrence of childhood sexual abuse, adult sexual assault, intimate partner violence, and sexual harassment: a mediational model of posttraumatic stress disorder and physical health outcomes. J Consult Clin Psychol. 2008 Apr;76(2):194-207.
[21] Bonomi AE, Anderson ML, Rivara FP, Thompson RS. Health outcomes in women with physical and sexual intimate partner violence exposure. J Womens Health (Larchmt). 2007 Sep;16(7):987-97.
[22] Golding JM. Intimate partner violence as a risk factor for mental disorders: a meta-analysis. J Fam Violence. 1999;14(2):99-132.
[23] Hegarty K, Gunn J, Chondros P, Small R. Association between depression and abuse by partners of women attending general practice: descriptive, cross sectional survey. Bmj. 2004 Mar 13;328(7440):621-4.
[24] Gilchrist G, Hegarty K, Chondros P, Herrman H, Gunn J. The association between intimate partner violence, alcohol and depression in family practice. BMC Fam Pract. 2010 Sept 27;11:72.
[25] Neria Y, Bromet EJ, Carlson GA, Naz B. Assaultive trauma and illness course in psychotic bipolar disorder: findings from the Suffolk county mental health project. Acta Psychiatr Scand. 2005 May;111(5):380-3.
[26] Rivara FP, Anderson ML, Fishman P, Bonomi AE, Reid RJ, Carrell D, et al. Intimate partner violence and health care costs and utilization for children living in the home. Pediatrics. 2007 Dec;120(6):1270-7.
[27] Moraes CL, Amorim AR, Reichenheim ME. Gestational weight gain differentials in the presence of intimate partner violence. Int J Gynaecol Obstet. 2006 Dec;95(3):254-60.
[28] Murphy CC, Schei B, Myhr TL, Du Mont J. Abuse: a risk factor for low birth weight? A systematic review and meta-analysis. Cmaj. 2001 May 29;164(11):1567-72.
[29] Rico E, Fenn B, Abramsky T, Watts C. Associations between maternal experiences of intimate partner violence and child nutrition and mortality: findings from Demographic and Health Surveys in Egypt, Honduras, Kenya, Malawi and Rwanda. J Epidemiol Community Health. 2010 Sep 14.

[30] Hasselmann MH, Reichenheim ME. Parental violence and the occurrence of severe and acute malnutrition in childhood. Paediatr Perinat Epidemiol. 2006 Jul;20(4):299-311.
[31] Smith J. What is the impact of intimate partner abuse on children? . In: Roberts G, Hegarty K, Feder G, eds. Intimate partner abuse and health professionals : new approaches to domestic violence. Edinburgh: Churchill Livingstone 2006:127-43.
[32] Humphreys J. Children of battered women. In: Campbell JH, J., ed. Nursing care of survivors of family violence. St Louis:Mosby 1993.
[33] Taket A, Wathen CN, Macmillan H. Should health professionals screen all women for domestic violence? PLoS Med. 2004 Oct;1(1):e4.
[34] Mulley K. Screening women for domestic violence. A rapid response letter BMJ 6 September 2002.
[35] Feder G. Responding to intimate partner violence: what role for general practice? Br J Gen Pract. 2006 Apr;56(525):243-4.
[36] Ramsay J, Rivas C, Feder G. Interventions to reduce violence and promote the physical and psychosocial wellbeing of women who experience partner violence: a systematic review of controlled evaluations London: Department of Health; 2005.
[37] Ramsay J, Carter Y, Davidson L, Dunne D, Eldridge S, Feder G, et al. Advocacy interventions to reduce or eliminate violence and promote the physical and psychosocial well-being of women who experience intimate partner abuse. Cochrane Database Syst Rev. 2009(3):CD005043.
[38] McFarlane JM, Groff JY, O'Brien JA, Watson K. Secondary prevention of intimate partner violence: a randomized controlled trial. Nurs Res. 2006 Jan-Feb;55(1):52-61.
[39] Tiwari A, Leung WC, Leung TW, Humphreys J, Parker B, Ho PC. A randomised controlled trial of empowerment training for Chinese abused pregnant women in Hong Kong. Bjog. 2005 Sep;112(9):1249-56.
[40] Feder G, Hester M, Williamson E, Dunne D. Reducing intimate partner violence. In: Trafton JA, William GP, eds. Best Practices in the behavioural management of health from pre-conception to adolescence. Los Altos, CA: Institute of Disease Management, 2008.
[41] Taft AJ, Small R, Hegarty KL, Lumley J, Watson LF, Gold L. MOSAIC (MOthers' Advocates In the Community): protocol and sample description of a cluster randomised trial of mentor mother support to reduce intimate partner violence among pregnant or recent mothers. BMC Public Health. 2009;9:159.
[42] UK National Screening Committee. What is screening? 2009 [cited November 2009]; Available from: URL: http://www.screening.nhs.uk/screening
[43] Andermann A, Blancquaert I, Beauchamp S, Déry V. Revisiting Wilson and Jungner in the genomic age: a review of screening criteria over the past 40 years. Bull World Health Organ. 2008 Apr;86(4):317-19.
[44] Waalen J, Goodwin MM, Spitz AM, Petersen R, Saltzman LE. Screening for intimate partner violence by health care providers. Barriers and interventions. Am J Prev Med. 2000 Nov;19(4):230-7.
[45] Randall T. AMA, joint commission urge physicians become part of solution to family violence epidemic. Jama. 1991 Nov 13;266(18):2524, 7.

[46] American Medical Association Diagnostic and Treatment Guidelines on Domestic Violence. Arch Fam Med. 1992 Sep;1(1):39-47.
[47] American Academy of Family Physicians. Family violence: an AAFP white paper. The AAFP Commission on Special Issues and Clinical Interests. Am Fam Physician. 1994 Dec;50(8):1636-40, 44-6.
[48] Emergency medicine and domestic violence. American College of Emergency Physicians. Ann Emerg Med. 1995 Mar;25(3):442-3.
[49] Randall T. ACOG renews domestic violence campaign, calls for changes in medical school curricula. Jama. 1992 Jun 17;267(23):3131.
[50] Nelson HD, Nygren P, McInerney Y, Klein J. Screening women and elderly adults for family and intimate partner violence: a review of the evidence for the U. S. Preventive Services Task Force. Ann Intern Med. 2004 Mar 2;140(5):387-96.
[51] Ramsay J, Richardson J, Carter YH, Davidson LL, Feder G. Should health professionals screen women for domestic violence? Systematic review. Bmj. 2002 Aug 10;325(7359):314.
[52] Wathen CN, MacMillan HL. Interventions for violence against women: scientific review. Jama. 2003 Feb 5;289(5):589-600.
[53] Department of Health. Responding to domestic abuse: a handbook for health professionals. London 2005.
[54] Family Violence Prevention Fund (U.S.). National consensus guidelines on identifying and responding to domestic violence victimization in health care settings. San Francisco, Calif.: Family Violence Prevention Fund 2002.
[55] Bradley F, Smith M, Long J, O'Dowd T. Reported frequency of domestic violence: cross sectional survey of women attending general practice. Bmj. 2002 Feb 2;324(7332):271.
[56] Ramsay J, Richardson J, Carter Y, Feder G. Appraisal of evidence about screening for domestic violence: National Screening Committee; 2001.
[57] Taft A, Broom DH, Legge D. General practitioner management of intimate partner abuse and the whole family: qualitative study. Bmj. 2004 Mar 13;328(7440):618.
[58] Zink T. Screening or Case finding Domestic Violence. Ann Fam Med 2007.
[59] Cole TB. Is domestic violence screening helpful? Jama. 2000 Aug 2;284(5):551-3.
[60] Ferris LE. Intimate partner violence. Bmj. 2004 Mar 13;328(7440):595-6.
[61] Sackett DL, Haynes RB, Guyatt G, Tugwell P. Clinical epidemiology : a basic science for clinical medicine. 2nd ed. Boston: Little, Brown 1991.
[62] Daniels JP, Khan KS. Chronic pelvic pain in women. Bmj. 2010 Oct 5;341:c4834.
[63] Daniels J, Gray R, Hills RK, Latthe P, Buckley L, Gupta J, et al. Laparoscopic uterosacral nerve ablation for alleviating chronic pelvic pain: a randomized controlled trial. Jama. 2009 Sep 2;302(9):955-61.
[64] Latthe P, Mignini L, Gray R, Hills R, Khan K. Factors predisposing women to chronic pelvic pain: systematic review. Bmj. 2006 Apr 1;332(7544):749-55.
[65] Anderson M, Smith L, Sidel V. What is social medicine? Mon Rev 2005;56(8):34 –48.
[66] WHO Commission on Social Determinants of Health. Closing the gap in a generation: health equity through action on the social determinants of health. Final report of the Commission on Social Determinants of Health. Geneva: WHO; 2008.

[67] Rose D, Trevillion K, Woodall A, Morgan C, Feder G, Howard L. Barriers and facilitators of disclosures of domestic violence by mental health service users: qualitative study. Br J Psychiatry. 2010 Dec 15 [Epub ahead of print].
[68] Kendrick T, Hegarty K, Glasziou P. Interpreting research findings to guide treatment in practice. Bmj. 2008;337:a1499.
[69] Westad C, McConnell D. Child Welfare Involvement of Mothers with Mental Health Issues. Community Ment Health J. 2011 Jan 18 [Epub ahead of print].
[70] McCauley J, Kern DE, Kolodner K, Dill L, Schroeder AF, DeChant HK. The "battering syndrome": prevalence and clinical characteristics of domestic violence in primary care internal medicine practices. Ann Intern Med. 1995 Nov 15;123(10):737 - 46.
[71] Hegarty K, Taft A, Feder G. Violence between intimate partners: working with the whole family. Bmj. 2008;337:a839.
[72] Feder GS, Hutson M, Ramsay J, Taket AR. Women exposed to intimate partner violence: expectations and experiences when they encounter health care professionals: a meta-analysis of qualitative studies. Arch Intern Med. 2006 Jan 9;166(1):22-37.
[73] Campbell J, Laughon K, Woods A. Impact of intimate partner abuse on physical and mental health: how does it present in clinical practice? In: Roberts G, Hegarty K, Feder G, eds. Intimate partner abuse and health professionals : new approaches to domestic violence. First ed. Edinburgh: Churchill Livingstone 2006:43-60.
[74] Wilson SN, van der Kolk B, Burbridge J, Fisler R, Kradin R. Phenotype of blood lymphocytes in PTSD suggests chronic immune activation. Psychosomatics. 1999 May-Jun;40(3):222-5.
[75] Inslicht SS, Marmar CR, Neylan TC, Metzler TJ, Hart SL, Otte C, et al. Increased cortisol in women with intimate partner violence-related posttraumatic stress disorder. Ann N Y Acad Sci. 2006 Jul;1071:428-9.
[76] Rabin BS. Stress, immune function, and health : the connection. New York: Wiley-Liss 1999.
[77] Dutton MA, Green BL, Kaltman SI, Roesch DM, Zeffiro TA, Krause ED. Intimate partner violence, PTSD, and adverse health outcomes. J Interpers Violence. 2006 Jul;21(7):955-68.
[78] Zink T, Elder N, Jacobson J, Klostermann B. Medical management of intimate partner violence considering the stages of change: precontemplation and contemplation. Ann Fam Med. 2004 May-Jun;2(3):231-9.
[79] Wathen CN, MacMillan HL, Rhodes K, Levinson W. Intervening in Abusive Relationships--Reply. 2003:2211-a-2.
[80] Gerbert B, Abercrombie P, Caspers N, Love C, Bronstone A. How health care providers help battered women: the survivor's perspective. Women Health. 1999;29(3):115-35.
[81] Streiner DL, Norman GR. Health measurement scales : a practical guide to their development and use. 3rd ed. Oxford ; New York: Oxford University Press 2003.
[82] Deeks JJ. Systematic reviews in health care: Systematic reviews of evaluations of diagnostic and screening tests. BMJ. 2001 July 21, 2001;323(7305):157-62.
[83] Jaeschke R, Guyatt GH, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help

me in caring for my patients? The Evidence-Based Medicine Working Group. Jama. 1994 Mar 2;271(9):703-7.
[84] Rose GA. Ischemic Heart Disease. Chest Pain Questionnaire. Milbank Mem Fund Q. 1965 Apr;43:32-9.
[85] Spitzer RL, Kroenke K, Williams JB. Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. Primary Care Evaluation of Mental Disorders. Patient Health Questionnaire. Jama. 1999 Nov 10;282(18):1737-44.
[86] Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med. 2001 Sep;16(9):606-13.
[87] Dietrich AJ, Oxman TE, Burns MR, Winchell CW, Chin T. Application of a depression management office system in community practice: a demonstration. J Am Board Fam Pract. 2003 Mar-Apr;16(2):107-14.
[88] Robins LN, Helzer JE, Croughan J, Ratcliff KS. National Institute of Mental Health Diagnostic Interview Schedule. Its history, characteristics, and validity. Arch Gen Psychiatry. 1981 Apr;38(4):381-9.
[89] O'Connor R. Measuring quality of life in health. Edinburgh ; New York: Churchill Livingston 2004.
[90] Pearsall J. The concise Oxford dictionary. 10th ed. / edited by Judy Pearsall. ed. Oxford: Oxford University Press 1999.
[91] Straus SE. Evidence-based medicine : how to practice and teach EBM. 3rd ed. / Sharon E. Straus ... [et al.]. ed. Edinburgh: Elsevier Churchill Livingstone 2005.
[92] Yerushalmy J. Statistical problems in assessing methods of medical diagnosis, with special reference to X-ray techniques. Public Health Rep. 1947;62:1432–49.
[93] Jaeschke R, Guyatt G, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid? Evidence-Based Medicine Working Group. Jama. 1994 Feb 2;271(5):389-91.
[94] Knottnerus JABe. The evidence base of clinical diagnosis. London: BMJ Books 2002.
[95] Sohal H. The sensitivity and specificity of four questions, (HARK) to identify intimate partner abuse in general practice. MSc thesis. London: University of London; 2005.
[96] Mayfield D, McLeod G, Hall P. The CAGE questionnaire: validation of a new alcoholism screening instrument. Am J Psychiatry. 1974 Oct;131(10):1121-3.
[97] Beresford TP, Blow FC, Hill E, Singer K, Lucey MR. Comparison of CAGE questionnaire and computer-assisted laboratory profiles in screening for covert alcoholism. Lancet. 1990 Aug 25;336(8713):482-5.
[98] Straus SE, McAlister FA, Sackett DL, Deeks JJ. The accuracy of patient history, wheezing, and laryngeal measurements in diagnosing obstructive airway disease. CARE-COAD1 Group. Clinical Assessment of the Reliability of the Examination-Chronic Obstructive Airways Disease. Jama. 2000 Apr 12;283(14):1853-7.
[99] Smith LF. The WOMB (Women's views of birth) antenatal satisfaction questionnaire: development, dimensions, internal reliability, and validity. Br J Gen Pract. 1999 Dec;49(449):971-5.
[100] Baker R, Preston C, Cheater F, Hearnshaw H. Measuring patients' attitudes to care across the primary/secondary interface: the development of the patient career diary. Qual Health Care. 1999 Sep;8(3):154-60.

[101] Hegarty K, Fracgp, Bush R, Sheehan M. The composite abuse scale: further development and assessment of reliability and validity of a multidimensional partner abuse measure in clinical settings. Violence Vict. 2005 Oct;20(5):529-47.
[102] Devins G. Psychiatric rating scales. Paper presented at the Clarke Institute of Psychiatry. Toronto, Ontario 1993.
[103] Streiner DL, Norman GR. Health measurement scales : a practical guide to their development and use. 4th ed. ed. Oxford: Oxford University Press 2008.
[104] NICE. Hypertension: management of hypertension in adults in primary care Clinical guideline; 2006 June 2006.
[105] Thompson MP, Basile KC, Hertz MF, Sitterle D. Measuring Intimate Partner Violence Victimization and Perpetration: A Compendium of Assessment Tools. Atlanta (GA): Centers for Disease Control and Prevention, National Center for Injury Prevention and Control; 2006.
[106] Goodwin LD. Changing conceptions of measurement validity: an update on the new standards. J Nurs Educ. 2002 Mar;41(3):100-6.
[107] Anastasi A. Evolving concepts of test validation. Annual Review of Psychology. 1986;37:1-15.
[108] Anastasi A. Psychological testing. 6th ed. New York: Macmillan ; London : Collier Macmillan 1988.
[109] Crocker L. Editorial: The great validity debate. Educational Measurement: Issues and Practice. 1997;16(2)(4).
[110] Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. Bmj. 2008 Apr 26;336(7650):924-6.
[111] International Organization for Standardization. International vocabulary of basic and general terms in metrology. Geneva: ISO 1993.
[112] Greenhalgh T. How to read a paper : the basics of evidence based medicine. London: BMJ 1997.
[113] Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982 Apr;143(1):29-36.
[114] Henson RK. Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. Measurement and Evaluation in Counseling and Development. 2001;34:177-89.
[115] Hegarty K, Sheehan M, Schonfeld C. A multidimensional definition of partner abuse: development and preliminary validation of the Composite Abuse Scale. J Fam Violence 1999;14(4):399-415.
[116] Nunnally JC. Psychometric theory. 2nd ed. ed. New York ; London: McGraw-Hill 1978.
[117] Kline P. A handbook of test construction : introduction to psychometric design. London: Methuen 1986.
[118] Aaronson N, Alonso J, Burnam A. Assessing health status and quality-of-life instruments: attributes and review criteria. Qual Life Res. 2002 May;11(3):193-205.
[119] Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20:37-46.
[120] Anderson JA. Point biserial correlation (insert sg20). Stata Tech Bull. 1994;17.

[121] Lenert L, Kaplan RM. Validity and interpretation of preference-based measures of health-related quality of life. Med Care. 2000 Sep;38(9 Suppl):II138-50.
[122] Bernal H, Wooley S, Schensul JJ. The challenge of using Likert-type scales with low-literate ethnic populations. Nurs Res. 1997 May-Jun;46(3):179-81.
[123] Krug EG, Mercy JA, Dahlberg LL, Zwi AB. The world report on violence and health. Lancet. 2002 Oct 5;360(9339):1083-8.
[124] Fontes LA. Ethics in family violence research: cross-cultural issues. Fam Relat. 1998 Jan;47(1):53-61.
[125] Lachs MS. Screening for family violence: what's an evidence-based doctor to do? Ann Intern Med. 2004 Mar 2;140(5):399-400.
[126] McLeod E, Bywaters P. Social work, health, and equality. London: Routledge 2000.
[127] Sen P, Humphreys C, Kelly L. Violence against women in the UK : CEDAW thematic shadow report 2003. London: Womankind Worldwide 2004.
[128] Humphreys C. A health inequalities perspective on violence against women. Health Soc Care Community. 2007 Mar;15(2):120-7.
[129] Gupta R, ed. From Homemakers to Jailbreakers: Southall Black Sisters. London: Zed Books 2003.
[130] Kasturirangan A, Krishnan S, Riger S. The impact of culture and minority status on women's experience of domestic violence. Trauma Violence Abuse. 2004 Oct;5(4):318-32.
[131] Guillemin F, Bombardier C, Beaton D. Cross-cultural adaptation of health-related quality of life measures: literature review and proposed guidelines. J Clin Epidemiol. 1993 Dec;46(12):1417-32.
[132] Mares P, Henley A, Baxter C. Health Care in Multiracial Britain. Cambridge: Health Education Council and National Extension College Trust Ltd. 1985.
[133] Hastrup K. Establishing an ethnicity. In: D. IP, ed. Semantic Anthropology ASA monograph. London: Academic Press, 1982:145-60.
[134] Bhopal RS. Ethnicity, race, and health in multicultural societies : foundations for better epidemiology, public health, and health care. Oxford ; New York: Oxford University Press 2007.
[135] Barth F. Ethnic groups and boundries: the social organization of cultural difference: Waveland Press 1998.
[136] Wolf ER. Europe and the people without history. University of California Press 1982:381.
[137] Kaplan JB, Bennett T. Use of race and ethnicity in biomedical publication. Jama. 2003 May 28;289(20):2709-16.
[138] Cohen AP. Symbolising boundaries: Identity and diversity in British cultures. Manchester: Manchester University Press 1986.
[139] Bhopal R. Is research into ethnicity and health racist, unsound, or important science? Bmj. 1997 Jun 14;314(7096):1751-6.
[140] Epstein S. Inclusion : the politics of difference in medical research. Chicago: University of Chicago Press 2007.
[141] Osborne NG, Feit MD. The use of race in medical research. Jama. 1992 Jan 8;267(2):275-9.

[142] Marks J. Human Biodiversity: Genes, Race and History (Foundations of human behaviour): Aldine Transaction 1995.
[143] American Academy of Pediatrics Committee on Pediatric Research. Race/ethnicity, gender, socioeconomic status - research exploring their effects on child health: A subject review. Pediatrics. 2000;105:1349-51.
[144] US Department of Health and Human Services. Mental health: culture, race and ethnicity - a supplement to mental health: A report to the surgeon general. 2001 17.12.08 [cited 9.01.09]; Available from: http://www.mentalhealth.org/cre/toc.asp
[145] American Psychological Association. Guidelines on Multicultural Education, Training, Research, Practice, and Organizational Change for Psychologists. Am Psychol. 2003 May;58(5):377-402.
[146] American Sociological Association. The importance of collecting data and doing social scientific research on race. Washington, DC; 2003.
[147] American Anthropological Association. Statement on "race." 1998, May 17 [cited December 16th, 2008]; Available from:
[148] Malley-Morrison K, Hines DA. Attending to the role of race/ethnicity in family violence research. J Interpers Violence. 2007 Aug;22(8):943-72.
[149] Baldwin JR. Redefining culture : perspectives across disciplines. Mahwah, N.J. ; London: Lawrence Erlbaum Associates 2006.
[150] Krieger N. Inequality, diversity, and health: thoughts on "race/ethnicity" and "gender". J Am Med Womens Assoc. 1996 Aug-Oct;51(4):133-6.
[151] Oppenheimer GM. Paradigm lost: race, ethnicity, and the search for a new population taxonomy. Am J Public Health. 2001 Jul;91(7):1049-55.
[152] Rodriguez MA, Saba G. Cultural competence and intimate partner abuse: health care interventions. In: Roberts G, Hegarty K, Feder G, eds. Intimate partner abuse and health professionals : new approaches to domestic violence. Edinburgh: Churchill Livingstone 2006:181-96.
[153] Senior PA, Bhopal R. Ethnicity as a variable in epidemiological research. Bmj. 1994 Jul 30;309(6950):327-30.
[154] Chapman RR, Berggren JR. Radical contextualization: contributions to an anthropology of racial/ethnic health disparities. Health (London). 2005 Apr;9(2):145-67.
[155] Pearson M. The politics of ethnic minority health studies London: Croom Helm 1986.
[156] Collins PH. The tie that binds: race, gender and US violence. Ethn Racial Stud. 1998;21(5):917 - 38.
[157] Richie BE. Foreword. In: Richie BE, Sokoloff NJ, Pratt C, eds. Domestic violence at the margins: Readings on race, class, gender and culture. New Brunswick, NJ: Rutgers University Press. 2005:xv-xviii.
[158] White A, ed. Social focus in brief: ethnicity 2002. London: Office for National Statistics 2002.
[159] Gilbert M. The holocaust : the Jewish tragedy. London: Fontana, 1987 1986.
[160] Gould SJ. The mismeasure of man. Rev. and expanded. ed. New York ; London: Norton 1996.
[161] BBC. Profile: Dr James Watson 2007 18.12.08 [cited 9.01.09]; Available from: http://news.bbc.co.uk/1/hi/uk/7051310.stm

[162] House ER, Haug C. Book Reviews: Riding The Bell Curve: A Review: Richard J. Herrnstein and Charles Murray New York: Free Press, 1994. 846 pp. 1995:263-72.
[163] Amnesty International. Racism and the administration of justice 2001 [cited 9.01.09]; Available from: http://www.amnesty.org/en/library/asset/ACT40/020/2001/en/dom-ACT400202001en.html#download
[164] Zubaran C. The Quest for Recognition: Brazilian Immigrants in the United States. Transcultural Psychiatry 2008 Dec;45(4):590-610.
[165] Ethnicity, race, and culture: guidelines for research, audit, and publication. Bmj. 1996 Apr 27;312(7038):1094.
[166] McKenzie K, Crowcroft NS. Describing race, ethnicity, and culture in medical research. Bmj. 1996 Apr 27;312(7038):1054.
[167] Campbell J, Jones AS, Dienemann J, Kub J, Schollenberger J, O'Campo P, et al. Intimate partner violence and physical health consequences. Arch Intern Med. 2002 May 27;162(10):1157-63.
[168] Btoush R, Campbell JC, Gebbie KM. Visits coded as intimate partner violence in emergency departments: characteristics of the individuals and the system as reported in a national survey of emergency departments. J Emerg Nurs. 2008 Oct;34(5):419-27.
[169] Hawkins SS, Lamb K, Cole TJ, Law C. Influence of moving to the UK on maternal health behaviours: prospective cohort study. Bmj. 2008 May 10;336(7652):1052-5.
[170] Krieger N, Fee E. Social class: the missing link in US health data. Int J Health Serv 1994;24:25-44.
[171] Wagner PJ, Mongan P, Hamrick D, Hendrick LK. Experience of abuse in primary care patients. Racial and rural differences. Arch Fam Med. 1995 Nov;4(11):956-62.
[172] McFarlane J, Parker B, Soeken K, Silva C, Reed S, McFarlane J, et al. Severity of abuse before and during pregnancy for African American, Hispanic, and Anglo women. J Nurse Midwifery. 1999 Mar-Apr;44(2):139-44.
[173] Sorenson SB, Telles CA. Self-reports of spousal violence in a Mexican-American and non-Hispanic white population. Violence Vict. 1991 Spring;6(1):3-15.
[174] Krishnan SP, Hilbert JC, VanLeeuwen D. Domestic violence and help-seeking behaviors among rural women: results from a shelter-based study. Fam Community Health. 2001 Apr;24(1):28-38.
[175] Walby S, Allen J. Domestic violence, sexual assault and stalking: findings from the British Crime Survey. London: Home Office Research, Development and Statistics Directorate; 2004. Report No.: Study 276.
[176] Dimmitt JH. Self-concept and woman abuse: a rural and cultural perspective. Issues Ment Health Nurs. 1995 Nov-Dec;16(6):567-81.
[177] Dimmitt J. Rural Mexican-American and non-Hispanic white women: effects of abuse on self-concept. J Cult Divers. 1995 Spring;2(2):54-63.
[178] Jones AS, Campbell, J.C., Schollenberger, J. Annual and lifetime prevalence of partner abuse in a sample of female HMO enrollees. Womens Health Issues. 1999;9:295-305.
[179] Dearwater SR, Coben JH, Campbell JC, Nah G, Glass N, McLoughlin E, et al. Prevalence of intimate partner abuse in women treated at community hospital emergency departments. Jama. 1998 Aug 5;280(5):433-8.

[180] Russo NF, Denious JE, Keita GP, Koss MP. Intimate violence and black women's health. Womens Health. 1997 Fall-Winter;3(3-4):315-48.
[181] Waltermaurer E, Watson CA, McNutt LA. Black women's health: the effect of perceived racism and intimate partner violence. Violence Against Women. 2006 Dec;12(12):1214-22.
[182] Raj A, Silverman JG. Immigrant South Asian women at greater risk for injury from intimate partner violence. Am J Public Health. 2003 Mar;93(3):435-7.
[183] Raj A, Silverman JG. Intimate partner violence against South Asian women in greater Boston. J Am Med Womens Assoc. 2002 Spring;57(2):111-4.
[184] Raj A, Silverman JG, McCleary-Sills J, Liu R. Immigration policies increase south Asian immigrant women's vulnerability to intimate partner violence. J Am Med Womens Assoc. 2005 Winter;60(1):26-32.
[185] Caetano R, Ramisetty-Mikler S, McGrath C. Acculturation, Drinking, and Intimate Partner Violence among Hispanic Couples in the United States: A Longitudinal Study. 2004:60-78.
[186] Caetano R, Ramisetty-Mikler S, Caetano Vaeth PA, Harris TR. Acculturation stress, drinking, and intimate partner violence among Hispanic couples in the U.S. J Interpers Violence. 2007 Nov;22(11):1431-47.
[187] Yoshihama M, Horrocks J. Posttraumatic stress symptoms and victimization among Japanese American women. J Consult Clin Psychol. 2002 Feb;70(1):205-15.
[188] Yoshihama M. Battered Women's Coping Strategies and Psychological Distress: Differences by Immigration Status. American Journal of Community Psychology. 2002;30(3):429-52.
[189] Kishor S, Johnson K, ORC Macro. MEASURE/DHS+ (Programme). Profiling domestic violence : a multi-country study. Calverton, MD: MEASURE DHS+, ORC Macro 2004.
[190] Saunders-Robinson MA. Battered women: an African American perspective. Abnf J. 1991 Fall;2(4):81-4.
[191] Torres S. A comparison of wife abuse between two cultures: perceptions, attitudes, nature, and extent. Issues Ment Health Nurs. 1991 Jan-Mar;12(1):113-31.
[192] Yoshihama M. Reinterpreting strength and safety in a socio-cultural context: dynamics of domestic violence and experiences of women of Japanese descent. . Children Youth Services Rev 2000;22:207-29.
[193] Yoshihama M. A web in the patriarchal clan system: tactics of intimate partners in the Japanese sociocultural context. Violence Against Women. 2005 Oct;11(10):1236-62.
[194] Niaz U. Violence against women in South Asian countries. Arch Womens Ment Health. 2003 Aug;6(3):173-84.
[195] Buchbinder E, Eisikovits Z. Battered women's entrapment in shame: a phenomenological study. Am J Orthopsychiatry. 2003 Oct;73(4):355-66.
[196] Hampton RL, Carillo R, Kim J. Domestic violence in African American communities. In: B.E. Richie, N. J. Sokoloff, C. Pratt, eds. Domestic violence at the margins : readings on race, class, gender, and culture. New Brunswick, NJ: Rutgers University Press 2005:127-41.
[197] Gill A. Patriarchal Violence in the Name of 'Honour.' International Journal of Criminal Justice Sciences. 2006 January;1(1).

[198] Chen PH, Rovi S, Vega M, Jacobs A, Johnson MS. Screening for domestic violence in a predominantly Hispanic clinical setting. Fam Pract. 2005 Dec;22(6):617-23.
[199] Sorenson SB. Violence against women. Examining ethnic differences and commonalities. Eval Rev. 1996 Apr;20(2):123-45.
[200] Sohal H, Eldridge S, Feder G. The sensitivity and specificity of four questions (HARK) to identify intimate partner violence: a diagnostic accuracy study in general practice. BMC Fam Pract. 2007;8:49.
[201] Lingard L, Albert M, Levinson W. Grounded theory, mixed methods, and action research. Bmj. 2008;337:a567.
[202] Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Med Res Methodol. 2003 Nov 10;3:25.
[203] Whiting PF, Weswood ME, Rutjes AW, Reitsma JB, Bossuyt PN, Kleijnen J. Evaluation of QUADAS, a tool for the quality assessment of diagnostic accuracy studies. BMC Med Res Methodol. 2006;6:9.
[204] Whiting P, Rutjes AW, Dinnes J, Reitsma J, Bossuyt PM, Kleijnen J. Development and validation of methods for assessing the quality of diagnostic accuracy studies. Health Technol Assess. 2004 Jun;8(25):iii, 1-234.
[205] Whiting P, Harbord R, Kleijnen J. No role for quality scores in systematic reviews of diagnostic accuracy studies. BMC Med Res Methodol. 2005;5:19.
[206] Canto JG, Kiefe CI, Williams OD, Barron HV, Rogers WJ. Comparison of outcomes research with clinical trials using preexisting data. Am J Cardiol. 1999 Oct 15;84(8):923-7, A6.
[207] Naish J, Sturdy R, Bobby J, Pereira F. The association between Asian ethnicity and prescribing rates in east London general practices: a database study. 1997:100-5.
[208] Peralta RL, Fleming MF. Screening for intimate partner violence in a primary care setting: the validity of "feeling safe at home" and prevalence results. J Am Board Fam Pract. 2003 Nov-Dec;16(6):525-32.
[209] Paranjape A, Rask K, Liebschutz J. Utility of STaT for the identification of recent intimate partner violence. J Natl Med Assoc. 2006 Oct;98(10):1663-9.
[210] Feldhaus KM, Koziol-McLain J, Amsbury HL, Norton IM, Lowenstein SR, Abbott JT. Accuracy of 3 brief screening questions for detecting partner violence in the emergency department.[see comment]. JAMA. 1997 May 7;277(17):1357-61.
[211] MacMillan HL, Wathen CN, Jamieson E, Boyle M, McNutt LA, Worster A, et al. Approaches to screening for intimate partner violence in health care settings: a randomized trial. JAMA. 2006 Aug 2;296(5):530-6.
[212] Tiwari A, Fong DY, Chan KL, Leung WC, Parker B, Ho PC. Identifying intimate partner violence: comparing the Chinese Abuse Assessment Screen with the Chinese Revised Conflict Tactics Scales. Bjog. 2007 Sep;114(9):1065-71.
[213] Reichenheim ME, Moraes CL. Comparison between the abuse assessment screen and the revised conflict tactics scales for measuring physical violence during pregnancy. J Epidemiol Community Health. 2004 Jun;58(6):523-7.
[214] Ernst AA, Weiss SJ, Cham E, Hall L, Nick TG. Detecting ongoing intimate partner violence in the emergency department using a simple 4-question screen: the OVAT. Violence Vict. 2004 Jun;19(3):375-84.

[215] Sherin KM, Sinacore JM, Li XQ, Zitter RE, Shakil A. HITS: a short domestic violence screening tool for use in a family practice setting. Fam Med. 1998 Jul-Aug;30(7):508-12.
[216] Zink T, Levin L, Putnam F, Beckstrom A. Accuracy of five domestic violence screening questions with nongraphic language. Clin Pediatr (Phila). 2007 Mar;46(2):127-34.
[217] Bonomi AE, Thompson RS, Anderson M, Rivara FP, Holt VL, Carrell D, et al. Ascertainment of intimate partner violence using two abuse measurement frameworks. Inj Prev. 2006 Apr;12(2):121-4.
[218] Coker AL, Pope BO, Smith PH, Sanderson M, Hussey JR. Assessment of clinical partner violence screening tools. J Am Med Womens Assoc. 2001;56(1):19-23.
[219] Brown JB, Lent B, Brett PJ, Sas G, Pederson LL. Development of the Woman Abuse Screening Tool for use in family practice. Fam Med. 1996 Jun;28(6):422-8.
[220] Brown JB, Lent B, Schmidt G, Sas G. Application of the Woman Abuse Screening Tool (WAST) and WAST-short in the family practice setting. J Fam Pract. 2000 Oct;49(10):896-903.
[221] Brown JB, Schmidt G, Lent B, Sas G, Lemelin J. [Screening for violence against women. Validation and feasibility studies of a French screening tool]. Can Fam Physician. 2001 May;47:988-95.
[222] Chen PH, Rovi S, Washington J, Jacobs A, Vega M, Pan KY, et al. Randomized comparison of 3 methods to screen for domestic violence in family practice. Ann Fam Med. 2007 Sep-Oct;5(5):430-5.
[223] Sagrestano LM, Rodriguez AC, Carroll D, Bieniarz A, Greenberg A, Castro L, et al. A comparison of standardized measures of psychosocial variables with single-item screening measures used in an urban obstetric clinic. J Obstet Gynecol Neonatal Nurs. 2002 Mar-Apr;31(2):147-55.
[224] McFarlane J, Parker B, Soeken K, Bullock L. Assessing for abuse during pregnancy. Severity and frequency of injuries and associated entry into prenatal care. Jama. 1992 Jun 17;267(23):3176-8.
[225] Connelly CD, Newton RR, Landsverk J, Aarons GA. Assessment of intimate partner violence among high-risk postpartum mothers: concordance of clinical measures. Women Health. 2000;31(1):21-37.
[226] Hudson WW, Mcintosh SR. The assessment of spouse abuse: Two quantifiable dimensions. J Marriage Fam. 1981;43:873-85.
[227] Straus MA, Hamby SL, Boney-McCoy S, Sugarman DB. The revised conflict tactics scales (CTS2): development and preliminary psychometric data. J Fam Issues. 1996;17:283-316.
[228] Chan KL. Study on child abuse and spouse battering: report on findings of household survey. [A consultancy study commissioned by the SWD of the HKSAR Government]. . Hong Kong: Department of social work & social administration. The University of Hong Kong; 2005.
[229] Gielen AC, O'Campo PJ, Faden RR, Kass NE, Xue X. Interpersonal conflict and physical violence during the childbearing year. Soc Sci Med. 1994 Sep;39(6):781-7.
[230] Stewart DE. Incidence of postpartum abuse in women with a history of abuse during pregnancy. Cmaj. 1994 Dec 1;151(11):1601-4.

[231] Wathen CN, MacMillan HL. Prevention of violence against women: recommendation statement from the Canadian Task Force on Preventive Health Care. Cmaj. 2003 Sep 16;169(6):582-4.
[232] Rabin RF, Jennings JM, Campbell JC, Bair-Merritt MH. Intimate partner violence screening tools: a systematic review. Am J Prev Med. 2009 May;36(5):439-45 e4.
[233] Yut-Lin W, Othman S. Early Detection and Prevention of Domestic Violence Using the Women Abuse Screening Tool (WAST) in Primary Health Care Clinics in Malaysia. Asia Pac J Public Health. 2008;20(2):102-16.
[234] Gregory A, Ramsay J, Agnew-Davies R, Baird K, Devine A, Dunne D, et al. Primary care Identification and Referral to Improve Safety of women experiencing domestic violence (IRIS): protocol for a pragmatic cluster randomised controlled trial. BMC Public Health. 2010;10(1):54.
[235] Fogarty CT, Brown JB. Screening for abuse in Spanish-speaking women. J Am Board Fam Pract. 2002 Mar-Apr;15(2):101-11.
[236] Dreyer G, Hull S, Aitken Z, Chesser A, Yaqoob MM. The effect of ethnicity on the prevalence of diabetes and associated chronic kidney disease. QJM. 2009;102(4):261-9.
[237] Rivas C. Negotiating psychological abuse: a qualitative study of white British, Caribbean and African women in inner London. PhD thesis. London: Queen Mary, London University; 2011.
[238] Nash ST. Through Black eyes: African American women's constructions of their experiences with intimate male partner violence. Violence Against Women. 2005 Nov;11(11):1420-40.
[239] Griffith EE, Young JL, Smith DL. An analysis of the therapeutic elements in a black church service. Hosp Community Psychiatry. 1984 May;35(5):464-9.
[240] Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D, et al. The Revised CONSORT Statement for Reporting Randomized Trials: Explanation and Elaboration. 2001:663-94.
[241] Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Bmj. 2003 Jan 4;326(7379):41-4.
[242] Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Clin Chem. 2003 Jan;49(1):7-18.
[243] Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Fam Pract. 2004 Feb;21(1):4-10.
[244] Lesser J. Negotiating national identity : immigrants, minorities, and the struggle for ethnicity in Brazil. Durham, N.C.: Duke University Press 1999.
[245] Schunemann H, Oxman A, Brozek J, Glasziou P, Jaeschke R, Vist G, et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. BMJ. 2008 May 17, 2008;336(7653):1106-10.
[246] Lynn MR. Determination and quantification of content validity. Nurs Res. 1986 Nov-Dec;35(6):382-5.
[247] Haynes RB, Sackett DL, Gordon HG, Tugwell P. Clinical epidemiology : how to do clinical practice research. 3rd ed. Philadelphia: Lippincott Williams & Wilkins 2006.

[248] Marmot MG. Evidence based policy or policy based evidence? Bmj. 2004 Apr 17;328(7445):906-7.

APPENDIX A:

Reference standards to identify IPV

Conflict Tactics Scale (CTS)

19 item scale assesses abuse within the past year. Seven point frequency scale (never, once, twice, 3-5 times, 6-10 times, 11-20 times, >20 times). Items range in severity from low in coerciveness to physical violence. Total scores vary from 15 to 105. Three subscales: verbal reasoning (three items, alpha= 0.69), verbal aggression (seven items, alpha= 0.84), and violence (nine items, alpha= 0.93).

Straus MA. Measuring Intrafamily Conflict and Violence: The Conflict Tactics (CT) Scales. Jounal of Marriage and The Family 1979;75-88.

Conflict Tactics Scale- Revised (CTS2)

The CTS2 is based on CTS. It includes more items for the three subscales and has two extra subscales (sexual coercion and physical injury from assault). Thus there is a total of 78 items representing five subscales. Each question is asked once for the respondent and for respondent's partner. "His/her" or "him/her" changed to "my partner." Greater distinction between minor and severe acts. Answering format simplified. Order of items mixed up. Good internal consistency across all five subscales (alphas for negotiation = 0.86; psychological aggression= 0.79; physical assault = 0.86; sexual coercion = 0.87 and injury 0.95).

Straus MA, Hamby SL, Boney-McCoy S, Sugarman DB. The Revised Conflict Tactics Scales (CTS2). Development and Preliminary Psychometric Data. Journal of Family Issues 1996; 17(3):283-316.

Index of Spousal Abuse (ISA)

Has 30 questions which can be administered in written or oral format. It assesses for physical (ISA-P, 11 items, alpha= 0.91) and non-physical (ISA-NP, 19 items, alpha= 0.93) abuse, using a Likert scale of one (never) to five (very frequently). Items are

weighted, summed and standardized for each scale. A complex weighted calculation required for final score. Scores range from 0 to 100. Higher scores represent more severe abuse. Cut off scores for non-physical abuse = 25, for physical abuse = 10.
Hudson WW, McIntosh SR. The Assessment of Spouse Abuse: Two Quantifiable Dimensions. Journal of Marriage and The Family 1981;873-888
Later a modified version of the ISA-P with 15 items developed for which Cronbach's alpha = 0.93.
Hudson WW. Partner Abuse Scale Physical. Tempe, Ariz: Walmyr Publishing; 1991.

Composite Abuse Scale (CAS)

The CAS consists of 30 items taken from the Conflict Tactics Scale, Measures of Wife Abuse, Inventory of Spouse Abuse and the Psychological Maltreatment of Women Inventory. It assesses physical, emotional and sexual abuse. It is composed of four dimensions of abuse: severe combined abuse, emotional abuse, physical abuse and harassment.

Dimensions and items of the Composite Abuse Scale:

Severe combined abuse:

Kept me from medical care
Used a knife or gun or other weapon
Locked me in the bedroom
Put foreign objects in my vagina
Refused to let me work outside the home
Raped me
Tried to rape me
Took my wallet and left me stranded

Emotional abuse:

Told me that I was crazy
Tried to convince family, friends and children that I was crazy
Became upset if dinner/housework wasn't done when they thought it should be

Told me that I wasn't good enough

Tried to keep me from seeing or talking to my family

Told me that I was stupid

Tried to turn my family, friends and children against me

Did not let me socialise with my female friends

Told me that I was ugly

Told me no one would ever want me

Blamed me for their violence

Physical abuse:

Shook me

Hit or tried to hit me with something

Pushed, grabbed or shoved me

Kicked me, bit me or hit with a fist

Slapped me

Threw me

Beat me up

Harassment:

Harassed me over the telephone

Harassed me at work

Followed me

Hung around outside my house

Hegarty K, Sheehan M, Schonfeld C. A multidimensional definition of partner abuse: Development and preliminary validation of the Composite Abuse Scale. Journal of Family Violence 1999; 14(4):399-415.

Hegarty K, Fracgp, Bush R, Sheehan M. The composite abuse scale: further development and assessment of reliability and validity of a multidimensional partner abuse measure in clinical settings. Violence Vict 2005; 20(5):529-547.

APPENDIX B:

BMC Family Practice

BioMed

Research article

Open Ac

The sensitivity and specificity of four questions (HARK) to identify intimate partner violence: a diagnostic accuracy study in general practice

Hardip Sohal, Sandra Eldridge and Gene Feder*

Address: Centre for Health Sciences, Barts and the London, Queen Mary's School of Medicine and Dentistry, 2 Newark Street, London, E1 2AT, UI

Email: Hardip Sohal - ahsohal@yahoo.co.uk; Sandra Eldridge - s.eldridge@qmul.ac.uk; Gene Feder* - g.s.feder@qmul.ac.uk

* Corresponding author

Published: 29 August 2007

BMC Family Practice 2007, **8**:49 doi:10.1186/1471-2296-8-49

Received: 8 December 2006

Accepted: 29 August 2007

This article is available from: http://www.biomedcentral.com/1471-2296/8/49

Abstract

Background: Intimate partner violence (IPV) including physical, sexual and emotional violence, causes short and long term ill-health. Brief questions that reliably identify women experiencing IPV who present in clinical settings are a pre-requisite for an appropriate response from health services to this substantial public health problem. We estimated the sensitivity and specificity of four questions (HARK) developed from the Abuse Assessment screen, compared to a 30-item abuse questionnaire, the Composite Abuse Scale (CAS).

Methods: We administered the four HARK questions and the CAS to women approached by two researchers in general practice waiting rooms in Newham, east London. Inclusions: women aged more than 17 years waiting to see a doctor or nurse, who had been in an intimate relationship in the last year. Exclusions: women who were accompanied by children over four years of age or another adult, too unwell to complete the questionnaires, unable to understand English or unable to give informed consent.

Results: Two hundred and thirty two women were recruited. The response rate was 54%. The prevalence of current intimate partner violence, within the last 12 months, using the CAS cut off score of ≥3, was 23% (95% C.I. 17% to 28%) with pre-test odds of 0.3 (95% C.I. 0.2 to 0.4). The receiver operator characteristic curve demonstrated that a HARK cut off score of ≥1 maximises the true positives whilst minimising the false positives. The sensitivity of the optimal HARK cut-off score of ≥1 was 81% (95% C.I. 69% to 90%), specificity 95% (95% C.I. 91% to 98%), positive predictive value 83% (95% C.I. 70% to 91%), negative predictive value 94% (95% C.I. 90% to 97%), likelihood ratio 16 (95% C.I. 8 to 31) and post-test odds 5.

Conclusion: The four HARK questions accurately identify women experiencing IPV in the past year and may help women disclose abuse in general practice. The HARK questions could be incorporated into the electronic medical record in primary care to prompt clinicians to ask about recent partner violence and to encourage disclosure by patients. Future research should test the effectiveness of HARK in clinical consultations.

Background

Violence against women is a global issue affecting millions who experience it and have to live with its consequences [1]. Intimate partner violence (IPV) including physical, sexual and emotional abuse is a major public health problem.

The WHO Violence Against Women study [2] found that the prevalence of lifetime physical violence and sexual violence by an intimate partner, among ever-partnered women varied from 15 to 71% in urban and rural settings in 10 countries. The prevalence of IPV is higher among women seeking primary care than in community surveys of the same geographic populations [3].

In a study in 12 east London general practices it was found that 41% of women waiting to see their general practitioner (GP) or practice nurse had experienced physical violence from a partner or former partner. 17% had experienced it within the past year [4].

IPV causes short and long term health problems. From controlled studies in a wide range of settings, we know that these include injury, chronic pain, gastrointestinal and gynaecological conditions (including sexually transmitted diseases) [5]. Consequences of IPV extend to perinatal health with it being an independent risk factor for deficit in gestational weight gain during pregnancy [6] and strong evidence of an IPV association with low birth weight [7].

The psychological health problems associated with domestic violence are no less serious and have psychological parallels with the trauma of being taken hostage and subjected to torture [8]. The most prevalent mental health sequelae of IPV are depression and post-traumatic stress disorder [9].

Women who have experienced physical or psychological violence are fifteen times more likely to abuse alcohol and nine times more likely to abuse drugs than are non-abused women, and there is evidence that substance abuse is a consequence as well as a potential cause of IPV [10]. Children exposed to domestic violence also often experience emotional and behavioural problems [11]. In the developing world it has been shown that children exposed to severe and recurrent IPV are more likely to be admitted with severe acute malnutrition [12].

It is difficult to calculate the exact societal economic impact of IPV but the costs are high. In the United States annual costs of intimate partner rape, physical assault, and stalking exceed $5.8 billion, nearly $4.1 billion of which is for direct medical and mental health care services [13]. In the United Kingdom the annual cost to the national health service of physical assaults is £1.2 billion [14].

The Department of Health in England now recommends that "All trusts should be working towards routine enquiry" [15]. In the US, the Family Violence Prevention fund consensus guidelines recommend that all adolescent and adult patients should be routinely asked about domestic violence [16]. Although there is ongoing debate about the evidence for screening or routine enquiry [17], there is unquestionably a need for clinicians to ask about domestic violence more often than they currently do.

A study of women attending general practices in east London found that only 17% of women experiencing IPV reported that their doctor had asked them about domestic violence [4]. We know that women who are experiencing violence want to disclose this to trusted doctors and get support [18], but that a high proportion of women who are experiencing abuse do not disclose this spontaneously in clinical consultations [4].

Short questions that reliably identify women experiencing IPV who present in clinical settings are a pre-requisite for developing an appropriate response from health services to this substantial public health problem [19].

Many primary health care professionals, including general practitioners (GPs) and practice nurses, occasionally enquire about domestic violence. It has not been adequately determined whether their questions identify women experiencing IPV.

Short tests

We have identified eleven short tools (see Additional file 1), for identification of women experiencing IPV [20-30]. Only three were validated in primary care settings [20-22]. The first study did not consider sexual abuse and had an unrepresentative sample: it was able to differentiate between self identified survivors of abuse and non-abused patients; there was no evidence that it was able to identify women who had experienced IPV in a general practice population [20]. The second reported no sensitivity or specificity; instead there was correlation between their tool and the reference test (Abuse Risk Inventory, $r = 0.69$, $p = 0.01$) but this does not necessarily indicate a valid and specific measure of IPV [21]. The third tool, a single question about safety, had low sensitivity, positive and negative predictive values (9%, 63% & 57% respectively) [22].

Outside of primary care settings, another two instruments did not consider sexual abuse [23,24]. One reported no sensitivity or specificity; only those who were positive on the index test were recruited into the study [23]. The sensitivity, specificity and positive predictive value of the

other test were too low for use by clinicians (65%, 80% & 51% respectively) [24].

The sixth tool [25] had a low positive predictive value: 56%. The seventh instrument started with an open question which makes it difficult to use as a standardised tool [26] and the eighth, Webster's "self-report check list," was not validated against an appropriate reference standard so there was no calculation of test indices [27]. Two further studies evaluated single item measures [28,29] and concluded that these may not be adequate in assessing for domestic violence.

We believe that the eleventh instrument, the AAS [30] has the most potential. Its strengths include that it covers a wide definition of partner violence which includes sexual abuse; a number of the aforementioned tools do not include sexual abuse [20,23,24,28]. It has 5 items rather than an unsatisfactory single item as is the case with a number of the tools [22,28,29]. Additionally it has a simple scoring system which we believe is important in brief general practice consultations unlike the likert scales used in 2 of the tools [20,25], the multiple scoring protocols in one [21] and an open question in another [26]. Finally, it has also been validated against an appropriate reference standard, the Index of Spouse Abuse (ISA) [31] unlike some [27].

However we also feel that the AAS has a number of weaknesses. Although the investigators concluded that the AAS questions were valid, this was based on a correlation between the score on a three-question version of the AAS and the ISA. No sensitivity or specificity was reported. Furthermore, the AAS validation was only within the setting of antenatal care in the US [30]. We do not know whether this is generalisable to other health care settings and in other countries, preventing its implementation into UK clinical practice [32].

More recently, in 2004, the test performance of the AAS was evaluated against the modified version of the conflict tactics scale (CTS 2) [33]. The AAS's sensitivity for minor physical violence was 32% and for severe physical violence was 61%. It was concluded that it was not sensible to use the AAS as a screening tool until more evidence was gathered.

In our study we have adapted the AAS, for use in a general practice setting, to form the HARK questionnaire (see table 1). We tested the HARK against the 30-item Composite Abuse Scale (CAS, see table 2) [34].

Table 1: HARK questions – one point is given for every yes answer

H	HUMILIATION Within the last year, have you been humiliated or emotionally abused in other ways by your partner or your ex-partner?
A	AFRAID Within the last year, have you been afraid of your partner or ex-partner?
R	RAPE Within the last year, have you been raped or forced to have any kind of sexual activity by your partner or ex-partner?
K	KICK Within the last year, have you been kicked, hit, slapped or otherwise physically hurt by your partner or ex-partner?

Methods

We conducted a cross-sectional survey of women in GP waiting rooms. The fifty-one general practices in Newham, a multi-ethnic inner city area of London, were stratified according to the number of doctors and the proportion of south Asian names on the practice register [35]. Equal numbers of practices were selected from each stratification group using a randomisation programme (SPSS version X). This was in an attempt to ensure that the practice population reflected the local area population.

Each practice was sent a recruitment letter with information about the study. If practices expressed an interest, a research team member met with the primary care team to answer any questions. We excluded practices that did not have a private room available, as then privacy for the survey could not be provided. If a practice decided not to take part or was excluded, the reason for this was documented and another practice was randomly selected from within the same stratification group.

We approached consecutive women in practice reception areas waiting to see a doctor or nurse. We included women aged more than 17 years who in the last year had been in an intimate relationship. We excluded women who were accompanied by children over four years of age or by another adult, were too unwell to complete the questionnaires, unable to understand English or unable to give informed consent. In the waiting room, women were asked to participate in a study designed to improve women's health care. We sought consent for the administration of the HARK and CAS questionnaires in a private room. All participants were given information on local domestic violence services. The East London and City ethics committee approved the study.

The number of potentially eligible subjects was recorded by the researcher in the waiting room. A record was made of the number of women who were excluded due to the exclusion criteria, those who the researchers were unable to approach at very busy times, women who were approached and agreed that they would be seen by the cli-

Table 2: Dimensions and items of the Composite Abuse Scale

Severe combined abuse	Kept me from medical care
	Used a knife or gun or other weapon
	Locked me in the bedroom
	Put foreign objects in my vagina
	Refused to let me work outside the home
	Raped me
	Tried to rape me
	Took my wallet and left me stranded
Emotional abuse	Told me that I was crazy
	Tried to convince family, friends and children that I was crazy
	Became upset if dinner/housework wasn't done when they thought it should be
	Told me that I wasn't good enough
	Tried to keep me from seeing or talking to my family
	Told me that I was stupid
	Tried to turn my family, friends and children against me
	Did not let me socialise with my female friends
	Told me that I was ugly
	Told me no one would ever want me
	Blamed me for their violence
Physical abuse	Shook me
	Hit or tried to hit me with something
	Pushed, grabbed or shoved me
	Kicked me, bit me or hit with a fist
	Slapped me
	Threw me
	Beat me up
Harassment	Harassed me over the telephone
	Harassed me at work
	Followed me
	Hung around outside my house

nician first and then undertake the study but were not seen again ("did not come back"), those who refused participation in the waiting room and those who declined consent in the private room.

The HARK and CAS were self-administered. We expected to be able to recruit approximately 500 women. On the assumption that the prevalence of IPV in the past year was 20%, we calculated that there was a 90% chance of estimating sensitivity at 76% or above with this sample size.

The Composite Abuse Scale – the reference standard

The CAS is a relatively robust standard for identifying IPV in primary care settings. It has an internal reliability (Cronbach's alpha) of .90 or more for each sub-scale, and all item-total score correlations of .6 or above [34]. It has also been validated with a large (1,836) sample of patients in general practice settings [36]. It is based on a concept of IPV that includes coercion, not simply violent acts arising out of conflict. It is recommended as an IPV research assessment tool by the National Centre for Injury Prevention and Control [37], as it has demonstrated reliability and validity for measuring the self-reported incidence and prevalence of IPV. It has evidence of content, construct, criterion and factorial validity. The CAS measures four dimensions of abuse inflicted on a woman by her partner: physical abuse (PA), emotional abuse (EA), severe combined abuse (SCA) and harassment. A preliminary cut-off score of 3 divides women presenting as abused or non-abused in general practice settings [36]. The 30 items are listed in table 2.

HARK – the index test

The acronym HARK denotes four short questions which represent different components of IPV. "Hark" is an archaic verb that means "to listen attentively." HARK arose out of an adaptation of the AAS. In HARK there is a focus only on IPV (not including that committed by a stranger), the pregnancy related item has been removed (so that it can be used in all women), for clarity emotional and physical violence are separated out into 2 items (rather than being combined in 1), "humiliation" was added (as it was thought to be plainer English and have a wider remit then "emotional abuse"), "rape" was added (to try to help cue a woman's memory by using language similar to her own) whilst items relating to fear and physical violence were directly retained from the AAS. The HARK questions are listed in table 1.

None of the women who were identified as having suffered abuse requested the researcher to make a direct referral in order to access specialised services.

Outcomes measures

The rate of current IPV within the last twelve months was calculated for the CAS (using the cut off score of ≥3) with 95% confidence intervals. This is equal to the prevalence or pre-test probability of IPV within the last twelve months.

The rates of IPV within the last twelve months were also calculated with 95% confidence intervals for the HARK, at different cut off scores (e.g. HARK cut off score ≥2, means a HARK score of either 2, 3 or 4). Each woman was identified as being positive or negative for IPV for each HARK cut off score and for the CAS cut off score of ≥3. We could then calculate HARK's sensitivity, specificity, positive predictive value (PPV – also known as the post-test probability), negative predictive value (NPV), likelihood ratios (LRs) with 95% confidence intervals and post-test odds (= pre-test odds × LR) at different HARK cut off scores [38].

A receiver operator characteristic (ROC) curve was constructed by plotting the sensitivity of each different HARK

cut off against the false positive rate (= 100 - specificity) at the different HARK cut offs. This was used to determine an optimal cut off HARK score which maximised the true positives whilst minimising the false positives.

Multilevel LRs [38] were also calculated at different HARK scores (e.g. a HARK score of 2 means 2 only, not ≥2, i.e. 2, 3 or 4) with 95% confidence intervals and corresponding post-test odds. 95% confidence intervals were calculated in EXCEL. Multilevel LRs allow exploration of the diagnostic usefulness of individual HARK scores.

We have used a variety of different methods to assess HARK's diagnostic accuracy at identifying IPV. Sensitivity and specificity interpret the HARK results retrospectively whereas PPVs and NPVs establish the predictive properties of the HARK in the future. The PPV is the proportion of women with a specific HARK result who are experiencing IPV. LRs express a result in terms of the actual chances of a woman experiencing IPV if her HARK score reaches a particular level. A LR for a given HARK result gives the odds that the test result comes from a person who is experiencing IPV. Unlike PPV and NPV, LRs are a good deal more constant with changes in prevalence. The post test odds allow background prevalence to be factored into the LR. Multilevel LRs express HARK's accuracy with level-specific likelihood ratios. They can be calculated at different HARK scores (e.g. 1) as opposed to cut offs (e.g. ≥1). They ensure that the maximum information is derived from the total range of possible HARK results (0 to 4).

Results

We approached 24 practices and 12 agreed to participate; 11 declined and one was excluded as it had no private room. Two hundred and thirty two women were recruited from May to October 2003. Figure 1 shows recruitment of individual participants to the study. Seven hundred and thirty seven women did not meet the inclusion criteria. Fourteen women were not approached because there were too many women in the waiting room for all to be approached. Two hundred and three women "did not come back." One hundred and eighty six women declined participation in the waiting room. Eleven women declined consent in the private room. The response rate of 54% (232/(232 + 186 + 11) was adjusted for the women who "did not come back."

The average age of participants was 35 years (range 18–70 years). 51% were in a paid job and 53% owned a house or flat. 40% of participants described their ethnic origin as white British, 25% as black British, African or Caribbean and 18% as Indian, Pakistani or Bangladeshi.

Outcomes measures

The CAS identified 53 cases of current IPV in the study population. This produced a prevalence (pre-test probability) of current IPV of 23% (95% C.I. 17% to 28%) with pre-test odds of 0.30 (95% C.I. 0.23 to 0.38). Pretest odds are prevalence divided by one minus prevalence.

Table 3 gives the sensitivity, specificity, PPV, NPV, LRs and post-test odds of HARK at different cut off scores. The receiver operator characteristic curve (figure 2) demonstrated that a HARK score ≥1 is the optimal cut off for detecting IPV. The predictive properties of the HARK score of ≥1 are highlighted in table 3. The HARK test accuracy (using a cut-off of ≥1) is 92%. This represents the proportion of true positives and true negatives as a proportion of all results.

Table 3: The sensitivity, specificity, PPV, NPV, LR & post-test odds with 95% confidence intervals of HARK at different cut off scores

Hark cut off scores	% of study sample	Sensitivity with 95% C.I.	Specificity with 95% C.I.	Positive predictive value with 95% C.I.	Negative predictive value with 95% C.I.	Likelihood ratio with 95% C.I.	Post-test odds
= 4	1%	4% (3% to 13%)	100% (98% to 100%)	100% (22 to 100%)	78% (72% to 83%)	Undefined	Undefined
≥3	6%	26% (15% to 40%)	100% (98% to 100%)	100% (81% to 100%)	82% (76% to 87%)	Undefined	Undefined
≥2	13%	51% (37% to 65%)	98% (95% to 100%)	90% (73% to 98%)	87% (82% to 91%)	30 (10 to 96)	9
≥1	22%	81% (69% to 90%)	95% (91% to 98%)	83% (70% to 91%)	94% (90% to 97%)	16 (8 to 31)	5
≥0	100%	100% (93% to 100%)	0% (0% to 2%)	23% (18% to 29%)	error	1	0.3

When the specificity is 100%, the likelihood ratio and post test odds are undefined.
Confidence intervals for likelihood ratios are approximate since they were calculated by the delta method which is less reliable when some cell sizes are small (Armitage P, Matthews JNS, Berry G. *Statistical methods in medical research*. Oxford; Blackwell, 1994).

Figure 1: Flow diagram to show recruitment of participants to the study

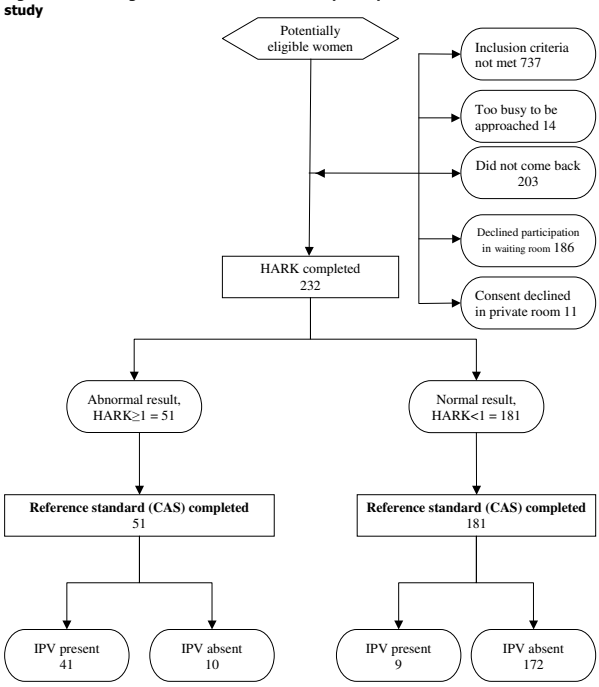


Figure 1
Flow diagram to show recruitment of participants to the study.

Multilevel LRs calculated at different HARK scores with 95% confidence intervals and corresponding post-test odds are shown in table 4.

Discussion

The four HARK questions accurately identify women experiencing IPV in the past year and may help women disclose IPV in general practice. The estimated specificity (95%, 95% C.I. 91% to 98%) of the HARK score of ≥1 was higher than the sensitivity (81%, 95% C.I. 69% to 90%). The PPV (post-test probabilities) of HARK, which increase as the HARK score increases, also provide evidence that HARK is an effective short tool for identifying IPV.

Table 4: Multilevel likelihood ratios with 95% confidence intervals and post-test odds of individual HARK scores.

HARK score (number of "yeses")	Likelihood ratio with 95% C.I.	Post-test odds
3 or 4	Undefined	Undefined
2	14.6 (4.3 to 49.4)	4.3
1	9.01 (3.7 to 21.9)	2.67
0	0.2 (0.1 to 0.4)	0.1

Figure 2: Receiver operator characteristic curve showing sensitivity of different HARK scores verses 1 – specificity.

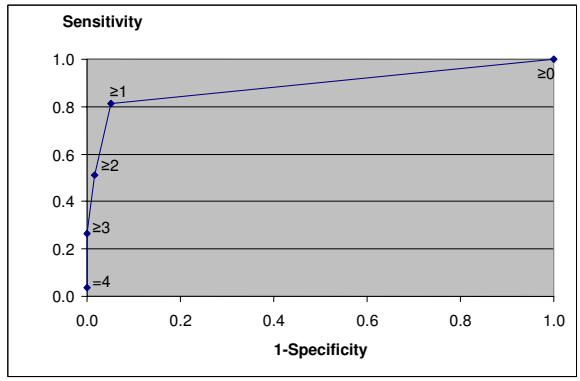


Figure 2
Receiver operator characteristic curve showing sensitivity of different HARK scores verses 1 - specificity.

The most straightforward way of using the HARK is as a simple test with a cut off of ≥1. Therefore if a clinician asks these four questions and their patient scores ≥1, this will identify 81% of women affected by IPV (as judged by the CAS). This is assuming that the tool performs in the same way that it did when a researcher administered it. There is an 83% probability that a woman with this score has experienced IPV in the past year (positive predictive value); and she is 16 times more likely to have been affected by IPV in the last year than some-one with a HARK score of 0 (likelihood ratio of a positive result).

The multilevel LRs and corresponding post-test odds make more use of the data from the test as it avoids dichotomising the HARK score into IPV present or not present [39]. When a woman is asked the four HARK questions she does not actually have a positive or negative score for IPV; instead she may score 0, 1, 2, 3 or 4 and each score has a different meaning (i.e. different likelihood ratio and post-test odds for IPV – see table 4). When an individual answers "no" to all of the HARK questions the likelihood ratio and post test odds (0.2 and 0.1 respectively) suggest that IPV is probably not present; whereas answering "yes" to three or four HARK questions produces a specificity of 100%, meaning that IPV is present. Answering "yes" to one or two of the HARK questions is less specific.

The majority of women who are experiencing IPV do not spontaneously disclose to clinicians. HARK can potentially accurately and quickly identify a high proportion of these women. This is a pre-requisite for effective intervention allowing the successful management of IPV in gen-

eral practice. It has been shown that women want to disclose IPV to health care professionals, particularly primary care clinicians [18].

The high pre-test probability (prevalence) of IPV (23%) is consistent with other prevalence studies in primary health care settings [3].

To increase the external validity of the study, we recruited a wide range of practices, including small single handed ones with less than 3,000 patients which are common in inner city areas in the United Kingdom. However small practices had fewer patients in the waiting room available for recruitment than had been anticipated; with the recruitment of participants taking longer than planned. Consequently we were only able to recruit 46% of our target sample size within the timeframe of the study, resulting in less precise estimates of test accuracy, reflected in wider confidence intervals. Nevertheless our study is larger than some other validation studies of short instruments and our estimates of test characteristics are relatively precise.

Eighty two percent of women who did not fulfil the inclusion criteria were accompanied. The ethics committee that approved our proposal specified that potential participants should only be approached if they were unaccompanied in order to decrease the likelihood of an abusive partner discovering that the participant had completed a questionnaire on domestic violence. We did include women who were accompanied by children under the age of five years, as it was felt that a child this young was unlikely to jeopardise a participant's safety.

Overall women were enthusiastic about participation once they found out that the study was about domestic violence: only eleven women declined consent in the private room. One hundred and eighty six women declined participation in the waiting room but these women did not know that the study was specifically about domestic violence.

The National Census 2001 figures allowed us to compare our study population to the local population in the borough of Newham. The average age of the study population was 3 years older than the average age in the local population (32 years). The percentage of the study population in a paid job was 12% higher and the percentage that owned a house or flat was 9% higher than that in the local population (39% and 44% respectively). The percentage of the study population that described their ethnic origin as white British was 6% higher than that in the local population (34%) whilst the percentage that described their ethnic origin as Indian, Pakistani or Bangladeshi was 11% lower than that in the local population (29%). This comparison shows that despite our attempts, the study population were not totally representative of the local population. We believe that the higher socio-economic status of our study sample (as reflected by the higher percentage in a paid job and owning a house or flat) compared to the local population may reflect a response bias meaning that perhaps those women with lower socio-economic status and at greater risk of IPV were less likely to have taken part in this study. This may have affected the calculation of the prevalence, PPV and NPV of HARK. However there is no reason why this would necessarily affect the sensitivity/specificity calculations unless the 46% of women who did not take part in the study answered differently with regards to only one of the instruments (the HARK or the CAS). This is unlikely.

The strengths of this study are that it tested a short tool that can be used in routine general practice, against an abuse measure validated in primary care. Additionally, HARK's external validity has been increased by being conducted in a range of practices with a study population of varied ethnicity.

Limitations included the response rate of 54%, decreasing the external validity of the study. Although we consider the CAS to be the best research measure for IPV in a health care setting, we cannot exclude the possibility that not all women who were found to be positive for IPV with the HARK but negative with the CAS were false positives. Other investigators have found that when using two sets of validated questions each may identify some women as abused that would have been missed by the other tool [40].

The HARK questions could be incorporated into electronic medical records in primary care to prompt clinicians to ask about recent intimate partner violence and to encourage disclosure by patients. Future research should test the effectiveness of HARK in clinical consultations as part of system level interventions to improve the response of primary care to IPV.

Conclusion

Intimate partner violence against women is common and causes short and long-term ill health. Previously questions about intimate partner violence to elicit disclosure have been insufficiently validated for use in general practice or family medicine populations, particularly outside the US. The four short HARK questions accurately identify women experiencing intimate partner violence in the past year.

Competing interests

The author(s) declare that they have no competing interests.

Authors' contributions

GF had the original idea of the study which was then developed by HS. SE advised on sample size and analysis. Guarantors: HS and GF.

All authors read and approved the final manuscript.

Additional material

Additional file 1

Table 1: Eleven short measures for detecting intimate partner violence in health care settings.

[http://www.biomedcentral.com/content/supplementary/1471-2296-8-49-S1.doc]

Acknowledgements

HS was supported by an East London and Essex research network (ELENoR) bursary. (The authors were independent from this funder). Thank you to Phillipa Chipping for practical advice on recruitment and Gurmit Kaur for helping collect data.

References

1. Watts C, Zimmerman C: **Violence against women: global scale and magnitude.** *Lancet* 2002, **359:**1232-1237.
2. **WHO multi-country study on women's health and domestic violence against women: summary report of initial results on prevalence, health outcomes and women's responses.** Geneva, World Health Organization; 2005.
3. Hegarty K: **What is intimate partner abuse and how common is it?** In *Intimate partner abuse and health professionals – new approaches to domestic violence* Edited by: Roberts G, Hegarty K, Feder G. London; Churchill Livingstone Elsevier; 2006:32-35.
4. Richardson J, Coid J, Petruckevitch A, Chung WS, Moorey S, Feder G: **Identifying domestic violence: cross sectional study in primary care.** *BMJ* 2002, **324:**274-277.
5. Campbell JC: **Health consequences of intimate partner violence.** *Lancet* 2002, **359:**1331-1335.
6. Moraes CL, Amorim AR, Reichenheim ME: **Gestational weight gain differentials in the presence of intimate partner violence.** *Int J Gynecol Obstetrics* 2006, **95(3):**254-260.
7. Murphy CC, Schei B, Myhr TL, Du Mont J: **Abuse: a risk factor for low birth weight? A systematic review and meta-analysis.** *Can Med Assoc* 2001, **4:**79-84.
8. Graham DLR, Rawlings E, Rimini N: **Survivors of terror: battered women, hostages and the stockholm syndrome.** In *Feminist perspectives on wife abuse* Edited by: Yllo K, Bograd M. London: Sage; 1998.
9. Golding JM: **Intimate partner violence as a risk factor for mental disorders: a meta-analysis.** *J Fam Violence* 2002, **14:**99-132.
10. Stark E, Flitcraft A: **Women at risk: domestic violence and women's health.** London: Sage; 1996.
11. Humphreys J: **Children of battered women.** In *Nursing care of survivors of family violence* Edited by: Campbell JC, Humphreys J. St Louis: Mosby; 1993.
12. Hasselmann MH, Reichenheim ME: **Parental violence and the occurrence of severe and acute malnutrition in childhood.** *Paediatr Perinat Epidemiol* 2006, **20:**299-311.
13. National Center for Injury Prevention and Control: *Costs of Intimate Partner Violence Against Women in the United States* Atlanta (GA): Centers for Disease Control and Prevention; 2003.
14. Walby S: *The cost of domestic violence* London: Department of Trade and Industry; 2004.
15. Department of Health: *Responding to domestic abuse: a handbook for health professionals, London* 2005.
16. Family Violence Prevention Fund: *National concensus guidelines on identifying and responding to domestic violence victimization in health care settings, San Francisco* 2004.
17. Ramsay J, Richardson J, Carter YH, Davidson LL, Feder G: **Should health professionals screen for domestic violence? Systematic review.** *BMJ* 2002, **325:**314-318.
18. Feder GS, Hutson M, Ramsay J, Taket AR: **Women exposed to intimate partner violence: expectations and experiences when they encounter health care professionals: a meta-analysis of qualitative studies.** *Arch Intern Med* 2006, **166(1):**22-37.
19. Nelson HD, Nygren P, McInerney Y, Klein J: **Screening women and elderly adults for family and intimate partner violence: a review of the evidence for the U.S. Preventitive Services Task Force.** *Ann Intern Med* 2004, **140:**387-96.
20. Sherin KM, Sinacore JM, Li X, Zitter RE, Shakil A: **HITS: A Short Domestic Violence screening Tool for Use in a Family Practice Setting.** *Fam Med* 1998, **30(7):**508-12.
21. Brown JB, Lent B, Sas G: **Application of the Woman Abuse Screening Tool (WAST) and WAST-Short in the Family Practice Setting.** *J Fam Pract* 2000, **49(10):**896-903.
22. Peralta RL, Fleming MF: **Screening for intimate partner violence in a primary care setting: the validity of "feeling safe at home" and prevalence results.** *J Am Board Fam Pract* 2003, **16(6):**526-532.
23. Heron SL, Thompson MP, Jackson E, Kaslow NJ: **Do responses to an intimate partner violence screen predict scores on a comprehensive measure of intimate partner violence in low-income black women?** *Ann Emerg Med* 2003, **42(4):**483-91.
24. Feldhaus KM, Kozio-Mclain J, Amsbury HL, Norton IM, Lowenstein SR, Abbott JT: **Accuracy of 3 brief screening questions for detecting partner violence in the emergency department.** *JAMA* 1997, **277(17):**1357-1361.
25. Ernst AA, Weiss SJ, Cham E, Hall L, Nick TG: **Detecting ongoing intimate partner violence in the emergency department using a simple 4 question screen: the OVAT.** *Violence Vict* 2004, **19(3):**375-384.
26. Reid AJ, Biringer A, Carroll JD, Midmer M, Wilson LM, Chalmers B, Stewart DE: **Using the ALPHA form in practice to assess antenatal psychosocial health.** *CMAJ* 1998, **159(6):**677-84.
27. Webster RN, Holt V: **Screening for partner violence: direct questioning or self report.** *Obstet Gynecol* 2004, **103:**299-303.
28. Sagrestano LM, Rodriguez AC, Carroll D, Bieniarz A, Geenberg A, Castro L, Nuwayhid B: **A comparison of standardised measures of psychosocial variables with single-item screening measures used in an urban obstetric clinic.** *J Obstet Gynecol Neonatal Nurs* 2002, **31:**147-155.
29. Connelly CD, Newton RR, Landsverk J, Aarons GA: **Assessment of intimate partner violence among high risk postpartum mothers: concordance of clinical measures.** *Women & Health* 2000, **31(1):**21-37.
30. McFarlane J, Parker B, Soeken K, Bullock L: **Assessing for abuse during pregnancy. Severity and frequency of injuries and associated entry into prenatal care.** *JAMA* 1992, **267:**3176-3178.
31. Hudson W, McIntosh S: **The assessment of spouse abuse: two quantifiable dimensions.** *J Marriage Fam* 1981:873-888.
32. Ferris LE: **Intimate partner violence.** *BMJ* 2004, **328:**595-6.
33. Reichenheim ME, Moraes CL: **Comparison between the abuse assessment screen and the revised conflict tactics scales for measuring physical violence during pregnancy.** *J Epidemiol Community Health* 2004, **58:**523-527.
34. Hegarty K, Sheehan M, Schonfeld C: **A multidimensional definition of partner abuse: development and preliminary validation of the Composite Abuse Scale.** *J Fam Violence* 1999, **14(4):**399-415.
35. Naish J, Sturdy P, Bobby J, Pereira F, Dolan S: **The association between Asian ethnicity and prescribing rates in east London: a database study.** *Health Informatics J* 1997, **3:**100-105.
36. Hegarty K, Bush R, Sheehan M: **The Composite Abuse Scale: further development and assessment of reliability and validity of a multidimensional partner abuse measure in clinical settings.** *Violence Vict* 2005, **20(5):**529-47.
37. Thompson MP, Basile KC, Hertz MF, Sitterle D: **Measuring Intimate Partner Violence Victimization and Perpetration: A Compendium of Assessment Tools.** Atlanta (GA): Centers for Disease Control and Prevention, National Center for Injury Prevention and Control; 2006.

38. Strauss SE, Richardson WS, Glasziou P, Haynes RB: *Evidence-Based Medicine. How to Practice and Teach EBM* Third edition. London; Elsevier Churchill Livingstone; 2005.
39. Deeks JJ, Altman DG: **Diagnostic tests 4: likelihood ratios.** *BMJ* 2004, **329:**168-9.
40. Bonomi AE, Thompson RS, Anderson M, Rivara FP, Holt VL, Carrell D, Martin DP: **Ascertainment of intimate partner violence using two abuse measurement frameworks.** *Inj Prev* 2006, **12:**121-124.

Pre-publication history

The pre-publication history for this paper can be accessed here:

http://www.biomedcentral.com/1471-2296/8/49/prepub

APPENDIX C:

Search strings

Cochrane Collaboration central register (CENTRAL/CCTR) search string

ID	Search
#1	**Battered Women** in **All Fields** in **all products**
#2	**Spouse Abuse** in **All Fields** in **all products**
#3	**Domestic Violence** in **All Fields** in **all products**
#4	**abuse* near/3 wom*n** in **Title, Abstract or Keywords** in **all products**
#5	**abuse* near/3 partner*** in **Title, Abstract or Keywords** in **all products**
#6	**abuse* near/3 spous*** in **Title, Abstract or Keywords** in **all products**
#7	**(wife or wives) near/3 batter*** in **Title, Abstract or Keywords** in **all products**
#8	**(wife or wives) near/3 abuse*** in **Title, Abstract or Keywords** in **all products**
#9	**partner* near/3 violen*** in **Title, Abstract or Keywords** in **all products**
#10	**spous* near/3 violen*** in **Title, Abstract or Keywords** in **all**

	products
#11	**dat* near/3 violen*** in **Title, Abstract or Keywords** in **all products**
#12	(**#1** OR **#2** OR **#3** OR **#4** OR **#5** OR **#6** OR **#7** OR **#8** OR **#9** OR **#10** OR **#11**)
#13	**Child Abuse** in **All Fields** in **all products**
#14	**Child Abuse, Sexual** in **All Fields** in **all products**
#15	(**#13** OR **#14**)
#16	(**#12** AND NOT **#15**)

Medline Search String

NOTE:

No. 58 is the total number of hits without mother$

No. 62 is the additional hits found with the inclusion of mother$

#	Search History
1	*Battered Women/
2	*Spouse Abuse/
3	*Domestic Violence/
4	(abuse$ adj3 wom#n).tw.
5	(abuse$ adj3 partner$).tw.
6	(abuse$ adj3 spous$).tw.
7	((wife or wives) adj3 batter$).tw.

8 ((wife or wives) adj3 abuse$).tw.

9 (partner$ adj3 violen$).tw.

10 (spous$ adj3 violen$).tw.

11 (dat$ adj3 violen$).tw.

12 or/1-11

13 *Child Abuse/

14 *Child Abuse, Sexual/

15 12 not (13 or 14)

16 *Women/

17 Female/

18 (wom#n or female$).tw.

19 *Adolescent/

20 (adolescen$ or teen$).tw.

21 or/16-20

22 screening.mp. or exp Mass Screening/

23 screen$.tw.

24 identif$.tw.

25 detect$.tw.

26 exp Diagnosis, Oral/ or exp Diagnosis/ or exp Nursing Diagnosis/ or diagnosis.mp.

27 diagnostic test.mp. or exp Diagnostic Tests, Routine/

28 medical history taking.mp. or exp Medical History Taking/

29 self disclosure.mp. or exp Self Disclosure/

30 (routine$ adj3 (ask$ or question$ or enquir$)).tw.

31 screening tool$.tw.

32 or/22-31

33 advocacy.mp. or exp Patient Advocacy/

34 exp Counseling/ or counsel$.mp.

35 mentor$.mp. or exp Mentors/

36 crisis intervention.mp. or exp Crisis Intervention/

37 risk assessment.mp. or exp Risk Assessment/

38 exp Social Welfare/

39 social support.mp. or exp Social Support/

40 help seeking.mp.

41 (information giving or giv$ information).mp.

42 (advice giving or giv$ advice).mp.

43 health behavior.mp. or exp Health Behavior/

44 patient education.mp. or exp Patient Education/

45 safety.mp. or exp Safety/

46 safety behav$.mp.

47 psychotherapy.mp. or exp Psychotherapy/

48 psychological therapy.mp.

49 problem solv$.mp. or exp Health Education/

50 self efficacy.mp. or exp Self Efficacy/

51 intervention.mp. or exp Intervention Studies/

52 evaluation.mp. or exp Evaluation Studies/

53 program evaluation.mp. or exp Program Evaluation/

54 or/33-53

55 15 and 32

56 15 and 21 and 54

57 55 or 56

58 mother$.mp. [mp=ti, ot, ab, nm, hw]

59 16 or 17 or 18 or 19 or 20 or 58

60 15 and 59 and 54

61 55 or 60

62 61 not 57

CINAHL search string

No.	Search term	
1	(BATTERED ADJ WOMEN).TI,AB.	
2	BATTERED-WOMEN.MJ. OR PARTNER-ABUSE.MJ. OR DOMESTIC-VIOLENCE.MJ. OR SPOUSE-ABUSE.MJ.	
3	(ABUSE$ NEAR (WOM$ OR PARTNER$ OR SPOUS$)).TI,AB.	
4	((WIFE OR WIVES) NEAR (BATTER$ OR ABUSE$)).TI,AB.	
5	(VIOLEN$ NEAR (PARTNER$ OR SPOUS$ OR DATE OR DATING)).TI,AB.	
6	1 OR 2 OR 3 OR 4 OR 5	
7	(CHILD ADJ ABUSE).TI,AB.	
8	CHILD-ABUSE.MJ. OR CHILD-ABUSE-SEXUAL.MJ.	
9	6 NOT (7 OR 8)	
10	(WOM$ OR FEMALE$).TI,AB.	
11	WOMEN.W..MJ.	
12	MOTHER$.TI,AB.	
13	MOTHERS.W..DE.	
14	(ADOLESCEN$ OR TEEN$).TI,AB.	
15	ADOLESCENT-HEALTH.MJ.	
16	10 OR 11 OR 12 OR 13 OR 14 OR 15	
17	SCREEN$.TI,AB.	
18	HEALTH-SCREENING#.DE.	
19	(IDENTIF$ OR DETECT$).TI,AB.	
20	DIAGNOS$3.TI,AB.	

21	**DIAGNOSIS#.W..DE.**	
22	**(DIAGNOSTIC ADJ TEST).TI,AB.**	
23	**CLINICAL-ASSESSMENT-TOOLS#.DE. OR DIAGNOSTIC-TESTS-ROUTINE#.DE. OR INSTRUMENT-VALIDATION#.DE.**	
24	**(MEDICAL ADJ HISTORY).TI,AB.**	
25	**PATIENT-HISTORY-TAKING#.DE.**	
26	**(PATIENT ADJ ASSESSMENT).TI,AB.**	
27	**PATIENT-ASSESSMENT#.DE.**	
28	**(SELF ADJ DISCLOSURE).TI,AB.**	
29	**SELF-DISCLOSURE#.DE.**	
30	**(ROUTINE NEAR (ASK$ OR QUESTION$ OR ENQUIR$)).TI,AB.**	
31	**(SCREENING ADJ TOOL).TI,AB.**	
32	**17 OR 18 OR 19 OR 20 OR 21 OR 22 OR 23 OR 24 OR 25 OR 26 OR 27 OR 28 OR 29 OR 30 OR 31**	
33	**ADVOCACY.TI,AB.**	
34	**PATIENT-ADVOCACY#.DE. OR CONSUMER-ADVOCACY#.DE.**	
35	**COUNSEL$.TI,AB.**	
36	**COUNSELING#.W..DE.**	
37	**(SOCIAL ADJ WORK).TI,AB.**	
38	**MENTOR$.TI,AB.**	
39	**MENTORSHIP#.W..DE.**	
40	**(CRISIS ADJ INTERVENTION).TI,AB.**	
41	**CRISIS-INTERVENTION#.DE.**	
42	**(RISK ADJ ASSESSMENT).TI,AB.**	
43	**RISK-ASSESSMENT#.DE.**	
44	**(SOCIAL ADJ WELFARE).TI,AB.**	
45	**SOCIAL-WELFARE#.DE.**	
46	**(SOCIAL ADJ SUPPORT).TI,AB.**	
47	**SUPPORT-PSYCHOSOCIAL#.DE. OR SOCIAL-NETWORKS#.DE.**	
48	**(SUPPORT ADJ GROUP$).TI,AB.**	
49	**SUPPORT-GROUPS#.DE.**	
50	**(HELP ADJ SEEKING).TI,AB.**	

51	**HELP-SEEKING-BEHAVIOR#.DE.**	
52	**(GIV$ NEAR (INFORMATION OR ADVICE)).TI,AB.**	
53	**PATIENT-EDUCATION#.DE. OR HEALTH-PROMOTION#.DE.**	
54	**(HEALTH ADJ BEHAVIO$).TI,AB.**	
55	**HEALTH-BEHAVIOR#.DE. OR HEALTH-EDUCATION#.DE.**	
56	**SAFETY.TI,AB.**	
57	**SAFETY#.W..DE. OR PATIENT-SAFETY#.DE.**	
58	**PSYCHOTHERAPY.TI,AB.**	
59	**PSYCHOTHERAPY#.W..DE.**	
60	**(PSYCHOLOGICAL ADJ THERAPY).TI,AB.**	
61	**(PROBLEM ADJ SOLV$).TI,AB.**	
62	**PROBLEM-SOLVING#.DE.**	
63	**(SELF ADJ EFFICACY).TI,AB.**	
64	**SELF-EFFICACY#.DE.**	
65	**(INTERVENTION$ OR EVALUATION$).TI,AB.**	
66	**INTERVENTION$.TI,AB.**	
67	**EVALUATION$.TI,AB.**	
68	**(PROGRAM ADJ EVALUATION).TI,AB.**	
69	**PROGRAM-EVALUATION#.DE.**	
70	**33 OR 34 OR 35 OR 36 OR 37 OR 38 OR 39 OR 40 OR 41 OR 42 OR 43 OR 44 OR 45 OR 46 OR 47 OR 48 OR 49 OR 50 OR 51 OR 52 OR 53 OR 54 OR 55 OR 56 OR 57 OR 58 OR 59 OR 60 OR 61 OR 62 OR 63 OR 64 OR 65 OR 66 OR 67 OR 68 OR 69**	
71	**9 AND 32**	
72	**9 AND 16 AND 70**	
73	**71 OR 72**	

BNID search string

No.	Database	Search term	
1	British Nursing Index - 1994 to date	(BATTERED ADJ WOMEN).TI,AB.	
2	British Nursing Index - 1994 to date	(SPOUSE ADJ ABUSE).TI,AB.	
3	British Nursing Index - 1994 to date	(DOMESTIC ADJ VIOLENCE).TI,AB.	
4	British Nursing Index - 1994 to date	DOMESTIC-VIOLENCE.DE.	
5	British Nursing Index - 1994 to date	(BATTER$ NEAR (WOM$ OR SPOUS$)).TI,AB.	
6	British Nursing Index - 1994 to date	(BATTER$ NEAR (PARTNER$ OR WIFE OR WIVES)).TI,AB.	
7	British Nursing Index - 1994 to date	(ABUS$ NEAR (WOM$ OR SPOUS$)).TI,AB.	
8	British Nursing Index - 1994 to date	(ABUS$ NEAR (PARTNER$ OR WIFE OR WIVES)).TI,AB.	
9	British Nursing Index - 1994 to date	(VIOLEN$ NEAR (WOM$ OR SPOUS$)).TI,AB.	
10	British Nursing Index - 1994 to date	(VIOLEN$ NEAR (PARTNER$ OR WIFE OR WIVES)).TI,AB.	
11	British Nursing Index - 1994 to date	(VIOLEN$ NEAR DAT$).TI,AB.	
12	British Nursing Index - 1994 to date	1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR 10 OR 11	
13	British Nursing Index - 1994 to date	(CHILD ADJ ABUSE).TI,AB.	
14	British Nursing Index - 1994 to date	CHILD-ABUSE-AND-NEGLECT.DE. OR CHILD-ABUSE-SEXUAL.DE.	

15	British Nursing Index - 1994 to date	**12 NOT (13 OR 14)**	

EMBASE search string

Search history: EMBASE

No.	Search term	
1	**BATTERED-WOMAN.MJ. OR PARTNER-VIOLENCE.MJ. OR DOMESTIC-VIOLENCE.MJ. OR FAMILY-VIOLENCE.MJ. OR BATTERING.W..MJ.**	
2	**(ABUSE$ NEAR WOM$).TI,AB.**	
3	**(ABUSE$ NEAR PARTNER$).TI,AB.**	
4	**(ABUSE$3 NEAR SPOUS$3).TI,AB.**	
5	**((WIFE OR WIVES) NEAR BATTER$).TI,AB.**	
6	**((WIFE OR WIVES) NEAR ABUSE$).TI,AB.**	
7	**(PARTNER$ NEAR VIOLEN$).TI,AB.**	
8	**(SPOUS$ NEAR VIOLEN$).TI,AB.**	
9	**(DAT$3 NEAR VIOLEN$3).TI,AB.**	
10	**1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9**	
11	**(CHILD ADJ ABUSE).TI,AB.**	
12	**CHILD-ABUSE.MJ.**	
13	**CHILD-SEXUAL-ABUSE.MJ.**	
14	**(CHILD$4 ADJ ABUSE ADJ SEXUAL).TI,AB.**	
15	**(CHILD$4 ADJ SEXUAL ADJ ABUSE).TI,AB.**	
16	**10 NOT (11 OR 12 OR 13 OR 14 OR 15)**	
17	**WOMEN.MJ.**	
18	**FEMALE.MJ.**	
19	**(WOM$3 OR FEMALE$3).TI,AB.**	
20	**(ADOLESCEN$ OR TEEN$).TI,AB.**	

21	**17 OR 18 OR 19 OR 20**	
22	**16 AND 21**	
23	**SCREENING-TEST#.DE.**	
24	**(MASS ADJ SCREENING).TI,AB.**	
25	**MASS-SCREENING#.DE.**	
26	**SCREEN$3.TI,AB.**	
27	**IDENTIF$.TI,AB.**	
28	**DETECT$3.TI,AB.**	
29	**DIAGNOSIS#.DE.**	
30	**(DIAGNOSTIC ADJ TEST).TI,AB.**	
31	**DIAGNOSTIC-TEST#.DE.**	
32	**(MEDICAL ADJ HISTORY ADJ TAKING).TI,AB.**	
33	**(SELF ADJ DISCLOSURE).TI,AB.**	
34	**SELF-DISCLOSURE#.DE.**	
35	**(ROUTINE$3 NEAR (ASK$3 OR QUESTION$5 OR ENQUIR$3)).TI,AB.**	
36	**(SCREENING ADJ TOOL$).TI,AB.**	
37	**SCREENING#.W..DE.**	
38	**23 OR 24 OR 25 OR 26 OR 27 OR 28 OR 29 OR 30 OR 31 OR 32 OR 33 OR 34 OR 35 OR 36 OR 37**	
39	**16 AND 38**	
40	**ADVOCACY.TI,AB.**	
41	**PATIENT-ADVOCACY#.W..DE.**	
42	**CONSUMER-ADVOCACY#.W..DE.**	
43	**PATIENT-COUNSELING#.W..DE.**	
44	**COUNSEL$.TI,AB.**	
45	**MENTOR$.TI,AB.**	
46	**(CRISIS ADJ INTERVENTION).TI,AB.**	
47	**CRISIS-INTERVENTION#.W..DE.**	
48	**(RISK ADJ ASSESSMENT).TI,AB.**	
49	**RISK-ASSESSMENT#.DE.**	
50	**(SOCIAL ADJ WELFARE).TI,AB.**	

51	**SOCIAL-WELFARE#.DE.**	
52	**(SOCIAL ADJ SUPPORT).TI,AB.**	
53	**SOCIAL-SUPPORT#.DE.**	
54	**(HELP ADJ SEEKING).TI,AB.**	
55	**HELP-SEEKING-BEHAVIOR#.DE.**	
56	**(INFORMATION ADJ GIVING).TI,AB.**	
57	**MEDICAL-INFORMATION#.DE.**	
58	**(GIV$3 ADJ INFORMATION).TI,AB.**	
59	**(ADVICE ADJ GIVING).TI,AB.**	
60	**(GIV$3 ADJ ADVICE).TI,AB.**	
61	**(HEALTH ADJ BEHAVIOR).TI,AB.**	
62	**HEALTH-BEHAVIOR#.DE.**	
63	**(PATIENT ADJ EDUCATION).TI,AB.**	
64	**PATIENT-EDUCATION#.DE. OR HEALTH-EDUCATION#.DE.**	
65	**SAFETY.TI,AB.**	
66	**SAFETY#.DE.**	
67	**PSYCHOTHERAPY.TI,AB.**	
68	**(PSYCHOLOGICAL ADJ THERAPY).TI,AB.**	
69	**(PROBLEM ADJ SOLV$3).TI,AB.**	
70	**PROBLEM-SOLVING#.DE.**	
71	**(HEALTH ADJ EDUCATION).TI,AB.**	
72	**(SELF ADJ EFFICACY).TI,AB.**	
73	**INTERVENTION.TI,AB.**	
74	**PATIENT-SAFETY#.DE.**	
75	**EVALUATION.TI,AB.**	
76	**PSYCHOTHERAPY#.DE.**	
77	**40 OR 41 OR 42 OR 43 OR 44 OR 45 OR 46 OR 47 OR 48 OR 49 OR 50 OR 51 OR 52 OR 53 OR 54 OR 55 OR 56 OR 57 OR 58 OR 59 OR 60 OR 61 OR 62 OR 63 OR 64 OR 65 OR 66 OR 67 OR 68 OR 69 OR 70 OR 71 OR 72 OR 73 OR 74 OR 75 OR 76**	
78	**22 AND 77**	
79	**39 OR 78**	

80	**MOTHER$.TI,AB.**	
81	**DIAGNOSIS.TI,AB.**	
82	**((MEDICAL OR PATIENT) ADJ HISTORY).TI,AB.**	
83	**21 OR 80**	
84	**38 OR 81 OR 82**	
85	**16 AND 84**	
86	**16 AND 83 AND 77**	
87	**85 OR 86**	

National research register search string

#1. ((batter* near woman) or (batter* near women) or (batter* near spouse) or (batter* near wife) or (batter* near wives) or (batter* near partner))

#2. ((abuse* near woman) or (abuse* near women) or (abuse* near spouse) or (abuse* near wife) or (abuse* near wives) or (abuse* near partner))

#3. ((violen* near woman) or (violen* near women) or (violen* near spouse) or (violen* near wife) or (violen* near wives) or (violen* near partner))

#4. ((violen* near dat*) or (domestic near violence) or (family near violence))

#5. ((child* near abuse*) or (child* near sex* near abuse*))

#6. (#1 or #2 or #3 or #4)

#7. (#6 and (not #5))

Top of Form

(#6 and (not #5)) - 169 hits

Save selected • Unselect all

- NRR Records from Regional and National Research Programmes
- NRR Records from Research Centres: Single-Centre Projects
- NRR Records from Research Centres: Lead Centres for Multi-Centre Projects
- NRR Records from Research Centres: Participating Centres for Multi-Centre Projects
- MRC Clinical Trials Directory
- CRD Register of Reviews (0 out of 276)
- Abstracts from The Cochrane Database of Systematic Reviews

Bottom of Form

Health Management Information Consortium search string

- #85 #83 not #84
- #84 (child* abuse)or(child* sexual abuse)
- #83 ((violen* near dat*)or(violen* near domestic)or(violen* near family)) or ((violen* near spous*)or(violen* near partner*)or(violen* near wi*)) or ((abuse* near partner*)or(abuse* near wi*)or(violen* near wom?n)) or ((batter* near wi*)or(abuse* near wom?n)or(abuse* near spous*)) or ((batter* near wom?n)or(batter* near spous*)or(batter* near partner*))
- #82 (violen* near dat*)or(violen* near domestic)or(violen* near family)
- #81 (violen* near spous*)or(violen* near partner*)or(violen* near wi*)
- #80 (abuse* near partner*)or(abuse* near wi*)or(violen* near wom?n)

- #79 (batter* near wi*)or(abuse* near wom?n)or(abuse* near spous*)
- #78 (batter* near wom?n)or(batter* near spous*)or(batter* near partner*)

Midwives Information and Resource Service (MIDIRS) search string

(batter* or abuse* or violen*) and (wom* or spous* or partner* or wife or wives or domestic or dating) and not (child* abuse or child* sexual abuse)

APPENDIX D:

Data collection form

Primary Reviewer		
Author (publication year)		
Title		
Ref Code		
Year data collected		
Tools used **Comparator:** **Index:**		
Sample used:	**N abused:**	**N non-abused:**
Original population		
Pre-enrolment exclusions		

Numbers eligible		
Numbers refusing		
Recruited		
Randomised		
Attrition		
Sample size calculation?		
Demographics		
Age: **Mean** **SD** **Range**		
Ethnicity		
Education		
Employed		
Income		
Marital status		
Mean yrs of marriage		
Currently		

pregnant		
Setting		
Inclusion criteria		
Exclusion criteria		
Tool presentation		
Whom presented tools		
Results: **Prevalence** **Reliability:** **1. Test Reliability** **2. Test-Retest** **3. Parallel Forms Coefficient**		

4. Internal Consistency Coefficient*	
5. Inter-Rater Reliability	
Validity:	
1. Content	
2. Criterion (Concurrent; Predictive)	
3. Construct (Convergent; Discriminant)	
2x2 Table	

Sensitivity **Specificity** **PPV** **NPV** **LR$_{pos}$** **LR$_{neg}$** **AUC** **DOR** **Any other statistical analyses computed:**	
Conclusions	
Strengths of paper (author)	
Limitations of paper (author)	

1st reviewer comments	

APPENDIX E:

Secondary data analyses

Tables showing diagnostic indices with 95% CIs and Figures showing Receiver operator characteristic curves

For south Asian groups (n = 48)

Description	Calculation of ratio	Calculated ratio	95% CI
Prevalence of IPV using CAS cut off score of ≥3	12/48	25.0%	0.1411 – 0.398946 = 14% to 40%
Pre-test odds	(12/48) / (36/48)	0.3333	

Hark Cut-off ≥1	Calculation of ratio	Calculated ratio	95% CI
Sensitivity	9/12	75.0%	0.428357 – 0.933064 = 43% to 93%
Specificity	35/36	97.2%	0.837965 - 0.998548 = 84% to 100%
Positive predictive value	9/10	90.0%	0.541155 - 0.994758 = 54% to 99%
Negative predictive value	35/38	92.1%	0.775159 - 0.979387 = 77% to 98%
Likelihood ratio	(9/12) / (1/36)	27	3.803329 - 191.674164 = 4 to 192
Post-test odds	(12/48) x(9/12) / (36/48) (1/36)	8.9999	

HARK POS ≥**1**	CASPOS 1	CASPOS 0
1	**9**	**1**
0	**3**	**35**

Hark Cut-off ≥0	Calculation of ratio	Calculated ratio	95% CI
Sensitivity	12/12	100.0%	0.698747 – 1 = 70% to 100%
Specificity	0/36	0.0%	0 - 0.120066 = 0% to 12%
Positive predictive value	12/48	25.0%	0.1411 - 0.398946 = 14% to 40%
Negative predictive value	0/0	Undefined	
Likelihood ratio	(12/12) / (36/36)	1.00	1 to 1
Post-test odds	(12/48) x(12/12) / (36/48) (36/36)	0.3333	

HARK POS ≥**0**	CASPOS 1	CASPOS 0
1	**12**	**36**
0	**0**	**0**

Hark Cut-off ≥2	Calculation of ratio	Calculated ratio	95% CI
Sensitivity	5/12	41.7%	0.164993 - 0.714007 = 16% to 71%
Specificity	35/36	97.2%	0.837965 - 0.998548 = 84% to 100%
Positive predictive value	5/6	83.3%	0.364823 - 0.991238 = 36% to 99%
Negative predictive value	35/42	83.3%	0.6804 - 0.92493 = 68% to 92%
Likelihood ratio	(5/12) / (1/36)	14.9999	1.940285 - 115.962319 = 2 to 116
Post-test odds	(12/48) x(5/12) / (36/48) (1/36)	4.9999	

HARK POS **≥2**	CASPOS 1	CASPOS 0
1	**5**	**1**
0	**7**	**35**

Hark Cut-off ≥3	Calculation of ratio	Calculated ratio	95% CI
Sensitivity	2/12	16.7%	0.029409 - 0.491185 = 3% to 49%
Specificity	36/36	100.0%	0.879934 – 1 = 88% to 100%
Positive predictive value	2/2	100.0%	0.197868 – 1 = 20% to 100%
Negative predictive value	36/46	78.3%	0.632407 - 0.885496 = 63% to 88%
Likelihood ratio	(2/12) / (0/36)	Undefined	
Post-test odds	(12/48) X (2/12) / (36/48)) (0/36)	Undefined	

HARK POS **≥3**	CASPOS 1	CASPOS 0
1	**2**	**0**
0	**10**	**36**

Hark Cut-off = 4	Calculation of ratio	Calculated ratio	95% CI
Sensitivity	0/12	0.0%	0 - 0.301253 = 0% to 30%
Specificity	36/36	100.0%	0.879934 – 1 = 88% to 100%
Positive predictive value	0/0	Undefined	
Negative predictive value	36/48	75.0%	0.601054 - 0.8589 = 60% to 86%
Likelihood ratio	(0/12) / (0/36)	Undefined	
Post-test odds	(12/48) X (0/12) / (36/48) (0/36)	Undefined	

HARK POS **≥4**	CASPOS 1	CASPOS 0
1	**0**	**0**
0	**12**	**36**

For African-Caribbean groups (n = 59)

Description	Calculation of ratio	Calculated ratio	95% CI
Prevalence of IPV using CAS cut off score of ≥3	19/59	32.20%	0.209686 - 0.457632 = 21% to 46%
Pre-test odds	(19/59) / (40/59)	0.4749	

Hark Cut-off ≥1	Calculation of ratio	Calculated ratio	95% CI
Sensitivity	17/19	89.5%	0.654618 - 0.981555 = 65% to 98%
Specificity	37/40	92.5%	0.785239 - 0.980428 = 78% to 98%
Positive predictive value	17/20	85.0%	0.611375 - 0.960434 = 61% to 96%
Negative predictive value	37/39	94.9%	0.813703 - 0.991068 = 81% to 99%
Likelihood ratio	(17/19) / (3/40)	11.93	3.974289 - 35.810357 = 4 to 36
Post-test odds	(19/59) x(17/19) / (40/59) (3/40)	5.6655	

HARK POS **≥1**	CASPOS 1	CASPOS 0
1	**17**	**3**
0	**2**	**37**

Hark Cut-off ≥0	Calculation of ratio	Calculated ratio	95% CI
Sensitivity	19/19	100.0%	0.790795 – 1 = 79% to 100%

Specificity	0/40	0.0%	0 - 0.109124 = 0 to 11%
Positive predictive value	19/59	32.2%	0.209686 - 0.457632 = 21% to 46%
Negative predictive value	0/0	Undefined	
Likelihood ratio	(19/19) (40/40)	1.00	1 to 1
Post-test odds	(19/59) x(19/19) (40/59) (40/40)	0.4749	

HARK POS ≥**0**	CASPOS 1	CASPOS 0
1	**19**	**40**
0	**0**	**0**

Hark Cut-off ≥2	Calculation of ratio	Calculated ratio	95% CI
Sensitivity	9/19	47.4%	0.25212 - 0.70505 = 25% to 70%
Specificity	40/40	100.0%	0.89088 – 1 = 89% to 100%
Positive predictive value	9/9	100.0%	0.62881 – 1 = 63% - 100%
Negative predictive value	40/50	80.0%	0.65856 - 0.89498 = 66% to 89%
Likelihood ratio	(9/19) (0/0)	Undefined	
Post-test odds	(19/59) x(9/19) (40/59) (0/0)	Undefined	

HARK POS **≥2**	CASPOS 1	CASPOS 0
1	**9**	**0**
0	**10**	**40**

Hark Cut-off ≥3	Calculation of ratio	Calculated ratio	95% CI
Sensitivity	6/19	31.6%	0.135554 - 0.565019 = 13% to 56%
Specificity	40/40	100.0%	0.890876 – 1 = 89% to 100%
Positive predictive value	6/6	100.0%	0.516818 – 1 = 52% to 100%
Negative predictive value	40/53	75.5%	0.614233 - 0.858096 = 61% to 86%
Likelihood ratio	(6/19) / (0/0)	Undefined	
Post-test odds	(19/59) x(6/19) / (40/59) (0/0)	Undefined	

HARK POS **≥3**	CASPOS 1	CASPOS 0
1	**6**	**0**
0	**13**	**40**

Hark Cut-off = 4	Calculation of ratio	Calculated ratio	95% CI
Sensitivity	1/19	5.3%	0.002754 - 0.281074 = 0% to 28%
Specificity	40/40	100.0%	0.890876 – 1 = 89% to 100%
Positive predictive value	1/1	100.0%	0.054621 – 1 = 5% to 100%
Negative predictive value	40/58	68.96%	0.553084 - 0.801021 = 55% to 80%
Likelihood ratio	(1/19) / (0/0)	Undefined	
Post-test odds	(19/59) x(1/19) / (40/59) (0/0)	Undefined	

HARK POS ≥**4**	CASPOS 1	CASPOS 0
1	**1**	**0**
0	**18**	**40**

For white groups (n = 112)

Description	Calculation of ratio	Calculated ratio	95% CI
Prevalence of IPV using CAS cut off score of ≥3	20/112	23%	0.11502 - 0.264745 = 11% to 26%
Pre-test odds	(20/112) / (92/112)	0.2174	

Hark Cut-off ≥1	Calculation of ratio	Calculated ratio	95% CI
Sensitivity	15/20	75.0%	0.505885 - 0.904067 = 51% to 90%
Specificity	88/92	95.652%	0.886193 - 0.985981 = 89% to 99%
Positive predictive value	15/19	78.9%	0.539021 - 0.930293 = 54% to 93%
Negative predictive value	88/93	94.6%	0.873243 - 0.980039 = 87% to 98%
Likelihood ratio	(15/20) / (4/92)	17.25	6.401518 - 46.483114 = 6 to 46
Post-test odds	(20/112) x(15/20) / (92/112) (4/92)	3.75015	

HARK POS**≥1**	CASPOS 1	CASPOS 0
1	**15**	**4**
0	**5**	**88**

Hark Cut-off ≥0	Calculation of ratio	Calculated ratio	95% CI
Sensitivity	20/20	100.0%	0.799547 – 1 = 80% to 100%
Specificity	0/92	0.0%	0 - 0.049947 = 0% to 5%
Positive predictive value	20/112	17.857%	0.11502 - 0.264745 = 11% to 26%
Negative predictive value	0/0	Undefined	
Likelihood ratio	(20/20) / (92/92)	1.00	1 to 1
Post-test odds	(20/112) x(20/20) / (92/112) (92/92)	0.2174	

HARK POS**≥0**	CASPOS 1	CASPOS 0
1	**20**	**92**
0	**0**	**0**

Hark Cut-off ≥2	Calculation of ratio	Calculated ratio	95% CI
Sensitivity	13/20	65.0%	0.40949 - 0.836913 = 41% to 84%
Specificity	90/92	97.8%	0.916209 - 0.996226 = 92% to 100%
Positive predictive value	13/15	86.7%	0.58389 - 0.976562 = 58% to 98%
Negative predictive value	90/97	92.8%	0.852018 - 0.968009 = 85% to 97%
Likelihood ratio	(13/20) / (2/92)	29.90	7.31469 - 122.221172 = 7 to 122
Post-test odds	(20/112) x(13/20) / (92/112) (2/92)	6.50	

HARK POS≥**2**	CASPOS 1	CASPOS 0
1	**13**	**2**
0	**7**	**90**

Hark Cut-off ≥3	Calculation of ratio	Calculated ratio	95% CI
Sensitivity	6/20	30.0%	0.128391 - 0.543307 = 13% to 54%
Specificity	92/92	100.0%	0.950053 – 1 = 95% to 100%
Positive predictive value	6/6	100.0%	0.516818 – 1 = 52% to 100%
Negative predictive value	92/106	86.8%	0.785 - 0.923301 = 79% to 92%
Likelihood ratio	(6/20) / (0/92)	Undefined*	
Post-test odds	(20/112) x(6/20) / (92/112) (0/92)	Undefined	

*VassarStats: defines this as infinity.

HARK POS≥**3**	CASPOS 1	CASPOS 0
1	**6**	**0**
0	**14**	**92**

Hark Cut-off = 4	Calculation of ratio	Calculated ratio	95% CI
Sensitivity	1/20	5.0%	0.002616 - 0.269443 = 0% to 27%
Specificity	92/92	100.0%	0.950053 – 1 = 95% to 100%
Positive predictive value	1/1	100.0%	0.054621 – 1 = 55% to 100%
Negative predictive value	92/111	82.9%	0.743036 - 0.89125 = 74% to 89%
Likelihood ratio	(1/20) / (0/92)	Undefined*	
Post-test odds	(20/112) x(1/20) / (92/112) (0/92)	Undefined	

*VassarStats: defines this as infinity.

HARK POS≥**4**	CASPOS 1	CASPOS 0
1	**1**	**0**
0	**19**	**92**

Comparing the three receiver operator characteristic curves in the African-Caribbean, south Asian and white groups:

ethnicity	Obs	ROC Area	-Asymptotic Normal-- Std. Err.	[95% Conf. Interval]	
1	48	0.8588	0.0669	0.72762	0.98998
2	59	0.9276	0.0394	0.85035	1.00000
3	112	0.8625	0.0514	0.76183	0.96317

Ho: area(1) = area(2) = area(3)

chi2(2) = 1.38 Prob>chi2 = 0.5022

This shows that there is no significant variation in the areas under the curves for these three different ethnic groups.

Individual HARK questions

Humiliation question

Description	Calculation of ratio	Calculated ratio	95% CI
Prevalence of IPV using CAS cut off score of ≥3	53/232	23%	17% to 28%
Pre-test odds	(53/232) / (179/232)	0.3	0.2 to 0.4

Humiliation question = 1	Calculation of ratio	Calculated ratio	95% CI
Sensitivity	37/53	69.8%	0.55488 - 0.81260 = 55% to 81%
Specificity	170/179	94.97%	0.90371 – 0.97527 = 90% to 97%
Positive predictive value	37/46	80.4%	0.65622 - 0.90138 = 66% to 90%
Negative predictive value	170/186	91.4%	0.86167 - 0.94844 = 86% to 95%
Likelihood ratio	(37/53) / (9/179)	13.88	7.1703 - 26.88657 = 7 to 27
Post-test odds	(53/232) x (37/53) / (179/232) (9/179)	4.16	

Humiliation question in south Asian groups

Humiliation question = 1	Calculation of ratio	Calculated ratio	95% CI
Sensitivity	7/12	58.3%	0.28599 - 0.83501 = 29% - 83%
Specificity	35/36	97.2%	0.83796 - 0.99855 = 84% - 100%
Positive predictive value	7/8	87.5%	0.46679 - 0.99344 = 47% - 99%
Negative predictive value	35/40	87.5%	0.72397 - 0.95305 = 72% - 95%
Likelihood ratio	(7/12) / (1/36)	21.00	2.86823 - 153.75337 = 3 to 154
Post-test odds	(12/48) x (7/12) / (36/48) (1/36)	6.99	

Humiliation question in African-Caribbean groups

Humiliation question = 1	Calculation of ratio	Calculated ratio	95% CI
Sensitivity	15/19	78.95%	0.53902 - 0.93029 = 54% to 93%
Specificity	37/40	92.5%	0.78524 - 0.98043 = 78% to 98%
Positive predictive value	15/18	83.33%	0.57735 - 0.95593 = 58% to 96%
Negative predictive value	37/41	90.2%	0.75941 - 0.96828 = 76% to 97%
Likelihood ratio	(15/19) / (3/40)	10.53	3.45929 to 32.03069 = 3 to 32
Post-test odds	(19/59) x (15/19) / (40/59) (3/40)	5.00	

Humiliation question in white groups

Humiliation question = 1	Calculation of ratio	Calculated ratio	95% CI
Sensitivity	13/20	65.0%	0.40949 - 0.83691 = 41% to 84%
Specificity	88/92	95.65%	0.88619 - 0.98598 = 89% to 99%
Positive predictive value	13/17	76.5%	0.49762 - 0.92177 = 50% to 92%
Negative predictive value	88/95	92.6%	0.84907 - 0.96733 = 85% to 97%
Likelihood ratio	(13/20) / (4/92)	14.95	5.43988 - 41.08594 = 5 to 41
Post-test odds	(20/112) x (13/20) / (92/112) (4/92)	3.25	

Afraid question

Afraid question = 1	Calculation of ratio	Calculated ratio	95% CI
Sensitivity	25/53	47.2%	0.33518 - 0.61230 = 33% to 61%
Specificity	176/179	98.3%	0.94789 - 0.99566 = 95% to 100%
Positive predictive value	25/28	89.3%	0.7063 - 0.97191 = 71% to 97%
Negative predictive value	176/204	86.3%	0.80603 - 0.90536 = 81% to 90%
Likelihood ratio	(25/53) / (3/179)	28.14	8.84355 - 89.57056 = 9 to 90
Post-test odds	(53/232) x (25/53) / (179/232) (3/179)	8.44	

Afraid question in south Asian groups

Afraid question = 1	Calculation of ratio	Calculated ratio	95% CI
Sensitivity	4/12	33.3%	0.11273 - 0.64563 = 11% to 65%
Specificity	35/36	97.2%	0.83796 - 0.99855 = 84% to 100%
Positive predictive value	4/5	80.0%	0.29879 - 0.98947 = 30% to 99%
Negative predictive value	35/43	81.4%	0.66082 - 0.91078 = 66% to 91%
Likelihood ratio	(4/12) / (1/36)	12.00	1.48181 - 97.17867 = 1 to 97
Post-test odds	(12/48) x (4/12) / (36/48) (1/36)	4.00	

Afraid question in African-Caribbean groups

Afraid question = 1	Calculation of ratio	Calculated ratio	95% CI
Sensitivity	9/19	47.4%	0.25212 - 0.70505 = 25% to 70%
Specificity	40/40	100.0%	0.89088 – 1 = 89% to 100
Positive predictive value	9/9	100.0%	0.62881 – 1 = 63% to 100%
Negative predictive value	40/50	80.0%	0.65856 - 0.89498 = 66% to 89%
Likelihood ratio	(9/19) / (0/40)	Undefined	
Post-test odds	(19/59) x (9/19) / (40/59) (0/40)	Undefined	

Afraid question in white groups

Afraid question = 1	Calculation of ratio	Calculated ratio	95% CI
Sensitivity	12/20	60.0%	0.36412 - 0.80022 = 36% to 80%
Specificity	90/92	97.8%	0.91621 - 0.99623 = 92% to 100%
Positive predictive value	12/14	85.7%	0.56151 - 0.97486 = 56% to 97%
Negative predictive value	90/98	91.8%	0.84084 - 0.96154 = 84% to 96%
Likelihood ratio	(12/20) / (2/92)	27.60	6.69341 - 113.80737 = 7 to 114
Post-test odds	(20/112) x (12/20) / (92/112) (2/92)	5.99	

Rape question

Only three women answered "yes" to this question; none were south Asian, two were African-Caribbean, one was white.

Rape question = 1	Calculation of ratio	Calculated ratio	95% CI
Sensitivity		5.7%	
Specificity		100.0%	
Positive predictive value	3/3	100.0%	
Negative predictive value	179/229	78.2%	
Likelihood ratio	(/) (/)		
Post-test odds	(/) x (/) (/) (/)		

Rape question in Asian groups

Rape question = 1	Calculation of ratio	Calculated ratio	95% CI
Sensitivity	0/12	0.0%	
Specificity	36/36	100.0%	
Positive predictive value	0/0	Undefined	
Negative predictive value	36/48	75.0%	
Likelihood ratio	(/) (/)	Undefined	
Post-test odds	(/) x (/) (/) (/)	Undefined	

Rape question in African-Caribbean groups

Rape question = 1	Calculation of ratio	Calculated ratio	95% CI
Sensitivity		10.5%	
Specificity		100.0%	
Positive predictive value	2/2	100.0%	
Negative predictive value	40/57	70.2%	
Likelihood ratio	(/) (/)	Undefined	
Post-test odds	(/) x (/) (/) (/)	Undefined	

Rape question in white groups

Rape question = 1	Calculation of ratio	Calculated ratio	95% CI
Sensitivity		5.0%	
Specificity		100.0%	
Positive predictive value	1/1	100.0%	
Negative predictive value	92/111	82.9%	
Likelihood ratio	(/) (/)	Undefined	
Post-test odds	(/) x (/) (/) (/)	Undefined	

Kick question

Kick question = 1	Calculation of ratio	Calculated ratio	95% CI
Sensitivity	21/53	39.6%	0.26760 - 0.53984 = 27% to 54%
Specificity	179/179	100.0%	0.973813 – 1 = 97% to 100%
Positive predictive value	21/21	100.0%	0.80760 – 1 = 81% to 100%
Negative predictive value	179/211	84.8%	0.79112 - 0.89250 = 79% to 89%
Likelihood ratio	(21/53) / (0/179)	Undefined	
Post-test odds	(53/232) x (21/53) / (179/232) (0/179)	Undefined	

Kick question in south Asian groups

Kick question = 1	Calculation of ratio	Calculated ratio	95% CI
Sensitivity	5/12	41.7%	0.16499 - 0.71401 = 16% to 71%
Specificity	36/36	100.%	0.71401 – 1 71% to 100%
Positive predictive value	5/5	100%	0.46294 – 1 = 46% to 100%
Negative predictive value	36/43	83.7%	0.68698 - 0.92672 = 69% to 93%
Likelihood ratio	(5/12) / (0/36)	Undefined	
Post-test odds	(12/48) x (5/12) / (36/48) (0/36)	Undefined	

Kick question in African-Caribbean groups

Kick question = 1	Calculation of ratio	Calculated ratio	95% CI
Sensitivity	7/19	36.8%	0.1723 - 0.61367 = 17% to 61%
Specificity	40/40	100.0%	0.89088 – 1 = 89% to 100%
Positive predictive value	7/7	100.0%	0.56093 – 1 = 56% to 100%
Negative predictive value	40/52	76.9%	0.62826 - 0.87019 = 63% to 87%
Likelihood ratio	(7/19) / (0/40)	Undefined	
Post-test odds	(19/59) x (7/19) / (40/59) (0/40)	Undefined	

Kick question in white groups

Kick question = 1	Calculation of ratio	Calculated ratio	95% CI
Sensitivity	9/20	45.0%	0.23829 - 0.67952 = 24% to 68%
Specificity	92/92	100.0%	0.95005 – 1 = 95% to 100%
Positive predictive value	9/9	100.0%	0.62881 – 1 = 63% to 100%
Negative predictive value	92/103	89.3%	0.81306 - 0.94288 = 81% to 94%
Likelihood ratio	(9/20) / (0/92)	Undefined	
Post-test odds	(20/112) x (9/20) / (92/112) (0/92)	Undefined	

Individual HARK questions and dimensions of IPV, as defined by CAS

Humiliation question (hark21) & Emotional IPV (eaposi)

Description	Calculation of ratio	Calculated ratio	95% CI
Prevalence of EA as defined by CAS EA Qs ≥3	47/232	20.3%	
Pre-test odds	(47/232) (185/232)	0.254	

Humiliation question = 1	Calculation of ratio	Calculated ratio	95% CI
Sensitivity	33/47	70.2%	0.54924 - 0.8221 = 55% to 82%
Specificity	172/185	92.9%	0.88024 - 0.9605 = 88% to 96%
Positive predictive value	33/46	71.7%	0.56319 - 0.83542 = 56% to 83%
Negative predictive value	172/186	92.4%	0.87442 - 0.95668 = 87% to 96%
Likelihood ratio	(33/47) (13/185)	9.99	5.72890 - 17.42679 = 6 to 17
Post-test odds	(47/232) x (33/47) (185/232) (13/185)	2.54	

Humiliation question (hark21) & EA (eaposi) in south Asian groups

Description	Calculation of ratio	Calculated ratio	95% CI
Prevalence of EA as defined by CAS EA Qs ≥3	12/48	25.0%	0.1411 - 0.39895 = 14% to 40%
Pre-test odds	(12/48) / (36/48)	0.3333	

Humiliation question = 1	Calculation of ratio	Calculated ratio	95% CI
Sensitivity	7/12	58.3%	0.28599 - 0.83500 =29% to 83%
Specificity	35/36	97.2%	0.83796 - 0.99855 = 84% to 100%
Positive predictive value	7/8	87.5%	0.46679 - 0.99344 = 47% to 99%
Negative predictive value	35/40	87.5%	0.72397 - 0.95305 = 72% to 95%
Likelihood ratio	(7/12) / (1/36)	20.9999	2.86823 - 153.75337 = 3 to 154
Post-test odds	(12/48) x (7/12) / (36/48) (1/36)	6.9999	

Humiliation question (hark21) & EA (eaposi) in African-Caribbean groups

Description	Calculation of ratio	Calculated ratio	95% CI
Prevalence of EA as defined by CAS EA Qs ≥3	15/59	25.4%	0.15373 - 0.38699 = 15% to 39%
Pre-test odds	(15/59) / (44/59)	0.341	

Humiliation question = 1	Calculation of ratio	Calculated ratio	95% CI
Sensitivity	12/15	80.0%	0.51373 - 0.94685 = 51% to 95%
Specificity	38/44	86.4%	0.71954 - 0.94332 = 72% to 94%
Positive predictive value	12/18	66.7%	0.41155 - 0.85643 = 41% to 86%
Negative predictive value	38/41	92.7%	0.78995 - 0.98091 = 79% to 98%
Likelihood ratio	(12/15) / (6/44)	5.87	2.67465 - 12.86813 = 3 to 13
Post-test odds	(15/59) x (12/15) / (44/59) (6/44)	2.00	

Humiliation question (hark21) & EA (eaposi) in white groups

Description	Calculation of ratio	Calculated ratio	95% CI
Prevalence of EA as defined by CAS EA Qs ≥3	19/112	17.0%	0.10775 - 0.25481 = 11% to 26%
Pre-test odds	(19/112) / (93/112)	0.20430	

Humiliation question = 1	Calculation of ratio	Calculated ratio	95% CI
Sensitivity	13/19	68.4%	0.43498 - 0.86445 = 43% to 86%
Specificity	89/93	95.7%	0.88736 - 0.98613 = 89% to 99%
Positive predictive value	13/17	76.5%	0.49762 - 0.92177 = 50% to 92%
Negative predictive value	89/95	93.7%	0.86228 - 0.97407 = 86% to 97%
Likelihood ratio	(13/19) / (4/93)	15.91	5.81622 - 43.50958 = 6 to 43
Post-test odds	(19/112) x (13/19) / (93/112) (4/93)	3.25	

Afraid question (hark22) & Emotional IPV (eaposi)

Description	Calculation of ratio	Calculated ratio	95% CI
Prevalence of EA as defined by CAS EA Qs ≥3	47/232	20.3%	
Pre-test odds	(47/232) (185/232)	0.254	

Afraid question = 1	Calculation of ratio	Calculated ratio	95% CI
Sensitivity	24/47	51.1%	0.36256 - 0.65699 = 36% to 66%
Specificity	181/185	97.8%	0.94199 - 0.99305 = 94% to 99%
Positive predictive value	24/28	85.7%	0.66438 - 0.95322 = 66% to 95%
Negative predictive value	181/204	88.7%	0.83371 - 0.92569 = 83% to 93%
Likelihood ratio	(24/47) (4/185)	23.62	8.61104 - 64.77305 = 9 to 65
Post-test odds	(47/232) x (24/47) (185/232) (41/185)	5.99	

Afraid question (hark22) & EA (eaposi) in south Asian groups

Description	Calculation of ratio	Calculated ratio	95% CI
Prevalence of EA as defined by CAS EA Qs ≥3	12/48	25.0%	0.1411 - 0.39895 = 14% to 40%
Pre-test odds	(12/48) / (36/48)	0.3333	

Afraid question = 1	Calculation of ratio	Calculated ratio	95% CI
Sensitivity	4/12	33.3%	0.11273 - 0.64563 = 11% to 65%
Specificity	35/36	97.2%	0.83796 - 0.99855 = 84% to 100%
Positive predictive value	4/5	80%	0.29879 - 0.98947 = 30% to 99%
Negative predictive value	35/43	81.4%	0.66082 - 0.91078 = 66% to 91%
Likelihood ratio	(4/12) / (1/36)	11.99	1.48181 - 97.17867 = 1 to 97
Post-test odds	(12/48) x (4/12) / (36/48) (1/36)	3.9999	

Afraid question (hark22) & EA (eaposi) in African-Caribbean groups

Description	Calculation of ratio	Calculated ratio	95% CI
Prevalence of EA as defined by CAS EA Qs ≥3	15/59	25.4%	0.15373 - 0.38699 = 15% to 39%
Pre-test odds	(15/59) / (44/59)	0.341	

Afraid question = 1	Calculation of ratio	Calculated ratio	95% CI
Sensitivity	9/15	60.0%	0.32891- 0.82543 = 33% to 82%
Specificity	44/44	100.0%	0.89999 - 1 = 90% to 100%
Positive predictive value	9/9	100.0%	0.62881 - 1 = 63% to 100%
Negative predictive value	44/50	88.0%	0.74997 - 0.95026 = 75% to 95%
Likelihood ratio	(9/15) / (0/44)	Undefined	
Post-test odds	(15/59) x (9/15) / (44/59) (0/44)	Undefined	

Afraid question (hark22) & EA (eaposi) in white groups

Description	Calculation of ratio	Calculated ratio	95% CI
Prevalence of EA as defined by CAS EA Qs ≥3	19/112	17.0%	0.10775 - 0.25481 = 11% to 26%
Pre-test odds	(19/112) / (93/112)	0.20430	

Afraid question = 1	Calculation of ratio	Calculated ratio	95% CI
Sensitivity	11/19	57.9%	0.33968 - 0.78879 = 34% to 79%
Specificity	90/93	96.8%	0.90192 - 0.991632 = 90% to 99%
Positive predictive value	11/14	78.6%	0.48816 - 0.94294 = 49% to 94%
Negative predictive value	90/98	91.8%	0.84084 - 0.96154 = 84% to 96%
Likelihood ratio	(11/19) / (3/93)	17.95	5.52933 - 58.25444 = 5 to 58
Post-test odds	(19/112) x (11/19) / (93/112) (3/93)	3.66	

Kick question (hark24) & Physical IPV (paposi)

Description	Calculation of ratio	Calculated ratio	95% CI
Prevalence of PA as defined by CAS PA Qs ≥1	41/232	17.7%	0.05823 - 0.13690 = 6% to 14%
Pre-test odds	(41/232) / (191/232)	0.21466	

Kick question = 1	Calculation of ratio	Calculated ratio	95% CI
Sensitivity	21/41	51.2%	0.35365 - 0.66849 = 35% to 67%
Specificity	191/191	100.0%	0.97542 - 1 = 97% to 100%
Positive predictive value	21/21	100.0%	0.80760 - 1 = 81% to 100%
Negative predictive value	191/211	90.5%	0.85539 - 0.93970 = 85% to 94%
Likelihood ratio	(21/41) / (0/191)	Undefined	
Post-test odds	(41/232) x (21/41) / (191/232) (0/191)	Undefined	

Kick question (hark24) & Physical IPV (paposi) in south Asian groups

Description	Calculation of ratio	Calculated ratio	95% CI
Prevalence of PA as defined by CAS PA Qs ≥1	9/48	18.7%	0.09438 - 0.33104 = 9% to 33%
Pre-test odds	(9/48) / (39/48)	0.23077	

Kick question = 1	Calculation of ratio	Calculated ratio	95% CI
Sensitivity	5/9	55.5%	0.22653 - 0.84657 = 23% to 85%
Specificity	39/39	100.0%	0.888332 - 1 = 89% to 100%
Positive predictive value	5/5	100%	0.46294 - 1 = 46% to 100%
Negative predictive value	39/43	90.7%	0.76946 - 0.96978 = 77% to 97%
Likelihood ratio	(5/9) / (0/39)	Undefined	
Post-test odds	(9/48) x (5/9) / (39/48) (0/39)	Undefined	

Kick question (hark24) & Physical IPV (paposi) in African-Caribbean groups

Description	Calculation of ratio	Calculated ratio	95% CI
Prevalence of PA as defined by CAS PA Qs ≥1	15/59	25.4%	0.15373 - 0.38699 = 15% to 39%
Pre-test odds	(15/59) / (44/59)	0.341	

Kick question = 1	Calculation of ratio	Calculated ratio	95% CI
Sensitivity	7/15	46.7%	0.22276 - 0.72577 = 22% to 73%
Specificity	39/39	100.0%	0.89999 - 1 = 90% to 100%
Positive predictive value	7/7	100.0%	0.56093 - 1 = 56% to 100%
Negative predictive value	44/52	84.6%	0.71367 - 0.92664 = 71% to 93%
Likelihood ratio	(7/15) / (0/39)	Undefined	
Post-test odds	(15/59) x (7/15) / (44/59) (0/39)	Undefined	

Kick question (hark24) & Physical IPV (paposi) in white groups

Description	Calculation of ratio	Calculated ratio	95% CI
Prevalence of PA as defined by CAS PA Qs ≥1	16/112	14.3%	0.08637 - 0.22461 = 9% to 22%
Pre-test odds	(16/112) / (96/112)	0.167	

Kick question = 1	Calculation of ratio	Calculated ratio	95% CI
Sensitivity	9/16	56.2%	0.30554 - 0.79246 = 30% to 79%
Specificity	96/96	100.0%	0.95205 - 1 = 95% to 100%
Positive predictive value	9/9	100.0%	0.62881 - 1 = 63% to 100%
Negative predictive value	90/98	93.2%	0.860218 - 0.96989 = 86% to 97%
Likelihood ratio	(9/16) / (0/96)	Undefined	
Post-test odds	(16/112) x (9/16) / (96/112) (0/96)	Undefined	

Note:

All values are calculated from the original 2x2 contingency table data in order to avoid rounding errors (i.e. LR is not calc by Sens/1-spec but, for example, by (43/53) / (9/179) as otherwise the rounding errors are considerable)