Assessing the use evidenced-informed care to reduce the overuse of medical interventions in intrapartum care: Validating the content of the Keeping Birth Normal Tool.

Florence Darling

City, University of London
Northampton Square
EC1V 0HB
London, United Kingdom
florence.darling@city.ac.uk
Telephone number: +4420889109876

317, St Margaret's Road,
Twickenham, Middlesex, TW1 1PN

Dr. Valentina Collington

St George's and Kingston University of London
Faculty of Health, Social Care and Education
Cranmer Terrace,
SW17 0RE
London, United Kingdom
v.collington@sgul.kingston.ac.uk

Abstract

Background: Unnecessary interventions in labour and birth increases the risk of mortality and morbidity in women. There are wide variation in the use of unnecessary interventions both regionally and globally. One of the reasons attributed to these variations is the poor implementation of evidence. This study validates the content of a new Tool to measure and support implementation where it is lacking.

Methods: Seven experts and eight women user representatives used a 4-point ordinal scale of relevance to rate fifty items in the Keeping Birth Normal Tool. Item-level content validity index (I-CVI), an average scale-level content validity index (S-CVI/Ave) and qualitative comments was used to delete and improve items.

Results: Eleven experts analysed all fifty items. Four experts rated thirty-five to forty-nine items. The initial scale received an S-CVI/Ave of 0.88. Two items were deleted, forty-five items improvement were made and seven new items added. The scale received an S-CVI/Ave of 1.0 post item deletion and improvement. Three further minor item improvements were made.

Discussion: The items in the KBN Tool are construct relevant. Future studies must gather evidence on response processes and internal structure to develop a Tool that is construct valid.

Introduction

Despite worldwide efforts to reduce cesarian-section rates, recently published data on global and regional trends show that it continues to rise (Betrán et al., 2015). Cesarian-section rates vary widely, in Europe for example a range of 14.8% to 52.2% is reported (Peri- Stat,). A rate of >15% is seen as medically unnecessary by the World Health Organisation (Gibbons et al, 2010). Reversing this trend, to control soaring healthcare costs, reduce maternal and neonatal mortality and specifically morbidities is necessary to long term health and well-being of women and babies (WHO, 2010; Betrán *et al*., 2015).

In the United Kingdom, ten years of evidenced-based policies and guidelines have been used to reduce unnecessary interventions and promote normal birth. An important recommendation is the use of midwife-led settings for women at low risk of complications (National Institute of Clinical Excellence, 2007). These settings include birth centres located in hospitals and community or care in the women's home. Midwives in these settings work autonomously seeking medical support when complications arise. This recommendation is supported by evidence in the UK and elsewhere that the use of unnecessary interventions are reduced in midwife-led settings (Brocklehurst et al., 2011; Hatem *et al.*, 2009*)*

However, women's choices are deeply influenced by views that obstetric unis are necessary to a safe birth (Coxon et al ., 2014) Care in these units are provided by obstetricians, midwives and other specialist. Many health care professional remain unconvinced about the safety of MLUs despite evidence. These views are unlikely to change very quickly. The obstetric unit remain an important choice for women and their families. Another important factor to consider within this context are variations in the use of unnecessary interventions. A recent audit demonstrated 1.5 to 2 - fold variation in the use of unnecessary interventions amongst

obstetric units in England. These variations do not appear to be influenced by social demographic factors or clinical risks (Royal College of Obstetrician and Gynecologist UK, 2014). Although the use of unnecessary interventions are lower amongst MLUs, use also varies for example, interventions are lower in birth centres located in the community as opposed to similar units in the hospital.

More research is needed to understand how the use of unnecessary interventions can be reduced in all environments for birth. This must include a greater understanding of mechanisms that lead to increased or reduced use of unnecessary interventions in order that the care women and babies receive is both equitable and of good quality (RCOG, 2014; Hollowell *et al*., 2015). Studies investigating the reasons for variations in outcomes amongst different birthing environments often cite medicalisation of birth, poor implementation of evidence and the lack of involvement of women in decision-making as obstructive to reducing unnecessary interventions (O' Connell and Downe, 2009; Walsh and Devane, 2010). The lack of midwifery skills and confidence to support a physiological birth has also been questioned (McCourt *et al*., 2012). This evidence is derived from small qualitative studies of variable quality. The inclusion of systematic measurement of care processes to support further qualitative work in efforts to produce more robust evidence has been proposed (Kennedy *et al*., 2009; Kings Fund, 2015; RCOG, 2015).

There is a paucity of Tools to measure skills that reduce unnecessary interventions. Many are focused on measuring technical aspects of care (Ramon *et al*., 2015). The newly developed Keeping Birth Normal Tool measures care under fifty items of evidenced-informed skills on reducing unnecessary interventions (Darling, 2016). A pilot showed that the Tool is useful and relevant to measuring and supporting implementation of evidenced- informed skills to

reduce unnecessary interventions (Darling, 2016). This supported the decision to validate the content of the Tool.

Validity and Content Validation

Current validation theory promotes the development of a Tool where the inferences based on the measurement using the Tool is valid. This referred to as construct validity (Messick, 1989, pp.13). This study uses an iterative process described by Kane (2013a) to develop construct validity. Two stages are described. In the development stage, different components of validity evidence are gathered. These include content, response process and internal structure and are dependent on the interpretation and use of the Tool. In the appraisal stage the Tool undergoes plausibility testing. Kane's model is endorsed by the Standards for Psychological and Educational Testing (AERA, APA and NCME, 1999).

The KBN Tool is in the developmental stage and this study gathers validity evidence about one component, its content. Content validity is an important component of construct validity because it provides evidence about the relevance of items in the Tool to the targeted construct (Lynn, 1985; Sireci and Bond, 2014). The targeted constructs that the KBN Tool measures are care processes that reduces unnecessary interventions and therapeutic alliance. Both these constructs are equally important in reducing unnecessary interventions in labour and first hour of birth (National Institute of Clinical Excellence, 2014; Walsh and Devane, 2012).

Methods

The design is pragmatic and employs mixed methods (Morgan, 2014). The content validation in this study uses judgement-quantification based on a standard developed through consensus (Lynn, 1985). The analysis uses a two stage process described in Lynn's (1985) seminal

work. It also draws Davis and Grant (1997), Haynes *et al.*, 1997, Polit *et al*. (2007) for guidance on:

- Selection and content experts sample size
- Selecting measures of inter-rater agreement
- Presentation of quantitative data
- Decision making on item deletion, improvement, additions
- Review procedure for the tool.

Quantitative data was derived from the assessment of relevance of the items. This data was used to make decision about item deletion. Traditional qualitative approaches uses of interviews or observations. In this study the experts were invited to provide written comments to improve the content. These comments were used to improve items (Lynn, 1985).

Sampling

This study uses a purposive sample of practitioners and women with known or demonstrable experience and expertise in normal birth (Grant and Davis, 1997). The clinical experts invited to participate had demonstrated clearly their pursuit and promotion of knowledge in the field of normal birth. This included teaching, practice, research, presentations and authoring of books or publication in peer reviewed journals (Grant and Davis, 1997). A total of n-15 were contacted via email

A pilot with community leads and their members was used to determine whether women could analyse content in a similar way to experts (Involve, 2012). The pilot showed that committee members who specifically engage users to improve quality of care could

participate. The lead representatives contacted a total of n = 12 via email. The researcher was copied into these emails.

The sample size used in judgement quantification is not based on calculations but dependent on the number of experts that can be identified, accessible, range of expertise and representation required. Grant and Davis (1997) propose three to twenty but Lynn (1985) states that at least five are necessary to control for chance agreements. A total of n=15 were recruited, 7 professionals and 8 women (Table 1).

Table 1: A Description of Content Experts

Experts		**Women**	
EP1	Midwife, research in normal birth locally and internationally working with the WHO and Europe, Academic. Widely published/Author	**EW1**	MSLC, AIMS, Revaluing Care network
EP2	Midwife, research into midwife –led care. Academic role on Knowledge Transfer. Widely published/Author	**EW2**	MSLC Chair, AIMS and Association of Breast-feeding Mothers
EP3	Midwife, research in normal birth, Chair of Nursing and Midwifery, widely published.	**EW3**	MSLC member, Doula
EP4	Midwife, Research Interest in patient-centred outcome, Organisational Systems and Cultures, Qualitative Methodology. Published.	**EW4**	MSLC Chair
EP5	Works with the Royal College of Midwives to promote midwife-led care. Widely published - indexed publications and author/editor of books.	**EW5**	MSLC member, Leads a birth choices group
EP6	Midwife, Researcher-PhD thesis on developing a tool, Education Project Manager. Widely published.	**EW6**	MSLC Chair
EP7	Midwife, lecturer, research interest-Art of Midwifery and Spirituality. Published.	**EW7**	Doula, Author
		EW8	MSLC Chair, Healthwatch Representative

Key: MSLC, UK = Maternity Services Liaison Committee , AIMS, UK= Association for the Improvements of Maternity Services

Data collection

Data Collection Tools

The instructions to the experts on the analysis of content is necessary to ensure a thorough and accurate analysis (Grant and Davis 1997). A pilot amongst two professional experts and two women tested the adequacy of the cover letter, Participation Information leaflet and Tool document. (Teijlingen and Hundley, 2001). The content in the Participation Information Leaflet was simplified to improve clarity. Sections in the Tool document were reworded to promote focus and understanding of the procedure. Research instruments were emailed.

Scale

A 4-point ordinal scale was used to analyse the items for relevance. This reduced bias from the use of a neutral option (Polit and Beck, 2006; Lynn, 1985). This was an important consideration in relation to the use of women in this study who may be reluctant to be critical (Teijlingen *et al*., 2003). A column for 'Not analysed' was introduced to allow for situations where women may not be able to assess item related to technical aspects of care. All participants were given time to consider participation in line with principles of good research practice (HRA, 2015). A reminder letter approved by the Ethics Committee was used to encourage response but avoid coercion. A maximum of two reminders were sent.

Content Validity index

The content validity index (CVI) is the measure of inter-rater agreement used in this study. The CVI defines the extent to which experts share a common interpretation of the constructs (Stemler, 2004). The Item – Level CVI and Scale-Level CVI were computed from items that received a score of three or four. The proportion of individual items given a relevant score generated an I-CVI. The proportion of items given a relevant rating by each expert generated

an S-CVI. The proportion of items given a relevant rating by all judges generated an S-CVI/Ave.

A modified Kappa statistic developed by Polit et al. 2007 (Table 2) was used to adjust each item-level CVI (I-CVI) for chance agreement of relevance. This was effective in identifying items for deletion and improvement (Polit *et al.*, 2007). A higher generalisable I-CVI standard of 0.78 was applied based on a standard developed by Polit *et al*. (2007) who demonstrated this as a safer generalisation regardless of the number of experts used.

Table 2: Evaluation criteria using the modified Kappa (k*)

I-CVI-adjusted for chance agreements (k*)	**Criteria for evaluation**
O.40-0.59	Fair
0.60-0.74	Good
>0.74	Excellent
>0.78	Excellent - Cut-off proposed by Polit et al., 2007, and used in this study

The S-CVI/Average was computed instead of the S-CVI/Universal agreement (S-CVI/UA) because the sample was large (Polit and Beck, 2006). The S-CVI/Ave also considers the risks of chance disagreements as well as non-chance disagreement in the event of bias or if construct specifications were misunderstood (Polit *et al*., 2007). A standard of 0.9 is used for SCI/Ave where the scale can be composed of some items with complete agreement and others with moderate agreement (Polit *et al*., 2007).

Ethics Approval

Ethical, legal and professional standards were guided by the Health Research Authority, Economic and Social Research Council and the Association of Internet Researchers (HRA, 2015, ESRC 2015, AoIR, 2012). Ethics approval was given the Research Ethics Committee

(REC) on the 20th May 2015. The local Research and Development approval was obtained on the 8th June.

Participation Information Leaflets including data collection instruments were sent to participants via email. All communications about the study with participants were retained as proof of communication. If completed Tool was returned this was accepted as implied consent. All principles of the Data Protection Act (1998) were adhered to in the management of data. Confidentiality and anonymity was ensured.

Data Analysis

Data was downloaded on an Excel spreadsheet. The Item-level CVI, S-CVI and S-CVI/AVE were calculated. Thematic analysis was used to manage participants' comments and improve items (Green and Thorogood, 2004, pp.177). The comments by the fifteen experts were collated under each item. These were compared and themes derived based on recurrent comments. These were cut and pasted under each theme derived under three categories (See table 6).

Results

Fifteen experts analysed the content. Eleven experts analysed all items. One expert did not analyse domain five. Two experts did not analyse two items in domain five. Two experts did not analyse two items in domain six. Two experts did not analyse four items in domain eight.

Two items, 5.1, 5.2 were deleted. Five other item improvements were made based on I-CVI scores (see Table3 and 4). A total of forty-five item improvements were made, forty with a

score of 0.78 and more. The reason for this was despite high I-CVIs values for individual items, only 47 % of the experts gave the overall scale, an S-CVI of 0.90. The average S-CVI/Ave was 0.88. The experts supported their ratings with qualitative comments. Seven new items were added. Three items remained unchanged.

Table 3: The I-CVI for each item

Item	I-CVI	Item	I-CVI	Item	I-CVI	Item	I-CVI	Item	I-CVI	Item	1-CVI
1.1	0.93	4.4	1.0	7.1	1.0	8.8	0.78	9.9	0.85	12.1	1.0
1.2	0.93	5.1	0.40	7.2	1.0	9.1	1.0	10.1	0.78	12.2	1.0
2.1	0.93	5.2	0.38	8.1	0.93	9.2	0.93	10.2	0.78	12.3	1.0
2.2	1.0	5.3	1.0	8.2	0.78	9.3	0.85	10.3	0.86	12.4	1.0
2.3	0.93	5.4	0.46	8.3	0.46	9.4	0.93	10.4	1.0	12.5	1.0
3.1	0.86	6.1	0.93	8.4	0.64	9.5	1.0	11.1	1.0		
4.1	0.93	6.2	0.93	8.5	0.86	9.6	1.0	11.2	1.0		
4.2	0.86	6.3	1.0	8.6	0.93	9.7	0.84	11.3	0.93		
4.3	0.93	6.4	0.85	8.7	0.64	9.8	0.64	11.4	1.0		

Pink -poor; Blue-fair; Green-Moderate; White-excellent

Table 4: Items for deletion and improvement based on I-CVI

Item	**I-CVI adjusted for chance agreements**	**Evaluation based on standards developed by Polit et al., 2007 using Fleiss (1981) and Cicchetti and Sparrow as a guide**	**Decision**
5.1	0.40	Poor	Item deleted
5.2	0.38	Poor	Item deleted
5.4	0.46	Fair	Item Improved
8.3	0.46	Fair	Item Improved
8.4	0.64	Good	Item Improved
8.7	0.64	Good	Item Improved
9.8	0.64	Good	Item Improved

Item Improvement based on qualitative comments

Five themes were identified under two categories that represented the constructs in the KBN Tool. Comments regarding language are addressed under a third category (See Table 5). Item improvements included merging of items, strengthening concepts, improving clarity and comprehensiveness.

Table 5: Themes derived from Qualitative Comments

Categories	**Themes**
1. Evidenced-informed-Therapeutic Alliance	Theme 1: Inclusion of women in decision-making
2. Evidenced-informed-Reducing overuse of medical Interventions	Theme 2: Surveillance increases risk of overuse
	Theme 2a: Quality of evidence 2.1a Evidence to support normal/physiological birth 2.2a. Evidence to support women in labour 2.3a. New evidence
3. Language	Theme 3a: Medicalised language Theme 3b: Clarity

There was a focus in expert comments on making the women the main decision-maker in their care. A need for flexibility and consideration of the woman's needs and perspectives were emphasized as opposed to only implementing evidence. Items were strengthen to reflect this.

Experts commented on the need for stronger conceptualization of care that reduces unnecessary interventions. For example "Need a descriptor that acknowledges that a physiological third stage is normative when preceded by a physiological 1st and 2nd stage" (EP2, EW1). The experts felt that support in labour in was not adequately represented and two new item were added. Items were also altered to capture the need to keep surveillance to the minimum and where there is a need to engage in it, to do so with as little disruption as

possible. This was so as to not disengage the women from normal hormonal physiology. Some items were updated to reflect current best evidence.

There were several comments on the use of language to empower or disempower to retain a focus on the surveillance repertoire of a medicalised birth. Wording were altered to reflect this concern. Six minor changes were also made to improve clarity. Where words were retained, justification is provided.

..

Calculation of I-CVI and S-CVI/Ave post review

Five expert verified the process of item deletion, improvement and made additional comments (Lynn, 1985; Haynes *et al*., 1997; Polit *et al*., 2007). The experts were selected on the basis of their capability evident in their analysis and critique. They were asked to analyse the revised content using a 4-point ordinal scale of relevance. I-CVI, S-CVI and S-CVI/Ave was calculated and a final rating provided (Polit *et al*, 2007).

All the items post review received an I-CVI score of 1.0 and an S-CVI/Ave of 1.0, an excellent rating. Three further minor improvements were made. Some of the items in the revised Tool are numbered differently as a result of revisions, merging of items and new items added. The five experts commented that the Tool was comprehensive measure of care to reduce unnecessary interventions in labour.

Discussion.

Fifteen experts analysed the relevance of items in the Tool to measure care to reduce unnecessary interventions in labour. Post analysis two items were deleted, forty-four items

were improved, seven new items added and three items remain unchanged. The Tool is now comprised of 12 domains and forty-five items and received an excellent rating on review by five experts. A weakness noted in studies on content validity are the lack of details about how indexes and qualitative comment were used to develop the scale (Polit et al., 2007; Lynn 1985). This study reports the I-CVI (Item-level content validity index) of individual items to demonstrate how items were selected for deletion and improvement. It used the SCV-I/Ave because of a large sample size.

The higher level of 0.78 for I-CVIs and S-CVI (scale-level content validity index) of 0.9 recommended by Polit et al. (2007) was used to develop good quality items. However the I-CVI identified only two items for deletion and five for improvement. This alone would not have resulted in the final excellent rating. Both S-CVI with a higher cut-off of 0.9 and qualitative comments were equally important in improving items.

Eight experts gave the initial scale an S-CVI of < 0.9. Despite being based on evidence, experts felt there was an inadequate focus on care processes to support physiological processes and involvement of women in decision-making. There was a level of confirmatory bias with several women giving the initial scale a score of 1. Women experts preferred to provide qualitative comments to improve items. A total of forty-five item improvements were made. The degree of improvements needed varied. Items with a universal agreement of 1.0 needed slight improvements in clarity. Qualitative data was used to improve items and was available for verification by the experts during review. The availability of this data allows for audit and replication and demonstrates a rigorous process to improve items (Morse *et al*., 2008).

Participants' comments were considered critically during item improvement. The study provided justification when decisions were made to retain items despite questioning of relevance by experts. An example is frequency of surveillance. These items received an I-CVI score of 0.38-0.65. However demonstrating safety is paramount in the practice environment (Kings Fund, 2008). The items were improved to include surveillance while minimising interference with physiological processes. This achieved a highly relevant rating on review.

Experts were also concerned that the involvement of women was not adequately considered in developing the scale. Aside from surveillance, items that measure the involvement of women in decision-making represent most of the improvements made. These are important changes in the current context of healthcare practice where involvement of women is evidenced as necessary to improving outcomes but is also rarely measured (Green, 2012; Greenhalgh, 2014). The inclusion of women as experts in this study strengthened the conceptualisation of items on involving women in care.

Some items were updated to reflect recent evidence. Tools need to be revised regularly to reflect current evidence to ensure that erroneous inferences are not made. At this point the Tool had undergone three iteration, and a wide range of improvements. The final analysis by five experts benefited from the use of participants who were drawn from the initial sample and conversant with the process. They had been critical but constructive in their initial analysis. This process of verification of quantitative data and the use of qualitative comments promoted the internal validity of the finding that items in the KBN Tool are construct relevant (Lynn, 1985).

Strengths and weakness

Collective expertise drawn from experience, constitutes a crucial element in the development of the KBN Tool. However, it inevitably introduces bias (Kane, 2013a). The iterative process described by Kane (2013a) can be used to address this. This will in the future include the gathering of evidence about response process, internal structure and plausibility testing.

Bias in the content validity stage was minimised by using a large sample of participants with a high level of expertise. This enabled a good control for chance agreement. Calculations of S-CVI included only rated items. The researcher remained separate from the process. The use quantification for item deletion and improvement also minimised researcher bias (Polit *et al*, 2007). Item improvements were based not only on quantification but qualitative comments. The use of comments used to improve items is available for audit and verification of rigor.

It is not possible to know if any communication took place between participants who are known to one another and so impacted on the ratings. The varied nature of the rating of individual items amongst at least eleven experts suggest that any impact from this is small. Consideration was given to non-respondents and impact on development. Experts were not questioned about non-participation, however, the quality of expertise of all participants was homogenous. On this basis it is unlikely non-participation would have made a difference to the development of the Tool.

The inclusion of women's in this study was given careful consideration. Guidance in validation literature about the use of expertise outside of clinical and academic experts is limited. Other sources of literature provided valuable information for developing a criteria to include women. Piloting played an important role in developing both the inclusion and

exclusion criteria and instruments for data collection. Confirmatory bias were minimised by giving women the opportunity to avoid analysing items that they felt unable to analyse. However confirmatory bias is unavoidable in content validation and the use of women may have increased bias. A future study may need to consider gathering validity evidence about response processes and internal structure depending on the interpretation that is to be derived from its use.

The piloting of the cover letter, Participation Information Leaflets and Tool document ensured that experts were clear about the constructs that each item under the domains were measuring and analysis of content using the scale of relevance. These elements including a careful selection of experts and an opportunity to provide qualitative comment ensured a thorough analysis, evident in the response of the experts.

The initial analysis by experts and several women was very critical of the items. After item improvements the Tool was given an excellent rating by five experts. Although the content of the Tool ultimately embraced most of the improvements recommended, a critical approach was used to retain items and improve items. The reality of practice in birth environment's and need to demonstrate safety was carefully considered.

This was an illuminating and humbling experience for the researcher who sees herself as an expert. It is evident from the initial ratings obtained and critique that she has been influenced to a large extent by a culture of surveillance and defensive practice. Similarly the pilot amongst practitioners who are influenced by a similar culture, resulted in minimal changes to the scale. The use of the judgement-quantification process highlighted these points resulting in the development of items that measures surveillance while minimising interruption of

physiological processes. Measures on involvement of women were expanded and strengthened.

Conclusions

This study develops a tool to measure a range of skills throughout labour and the first hour of birth to reduce unnecessary interventions. A recognized standard was used to inform the content validation. Quantification strengthened findings and measures of inter-rater agreement that were used have been widely tested and evaluated. The use of the content validity index as the measure of inter-rater agreement is justified in a study where content is in the early stages of development and decisions about item deletion and improvement needed to be made. The use of a generalisable standard for the I-CVI and S-CVI derived from controlling for chance agreements on relevant items promoted the validity of decision-making for deleting and improving items. The use of higher levels of I-CVI, S-CVI and qualitative comments contributed to the development of good quality items.

After the initial analysis two items were deleted and forty-five item were improved. Seven new items were added. The analysis of items post-improvement resulted in an excellent rating of the scale. Three further minor item improvements were made. The scale is comprised of twelve domains and forty-five items to measure evidence to reduce unnecessary interventions in intrapartum care and the involvement of women in decision-making.

The entire process was enhanced by careful selection of experts and piloting of instructions to ensure effective data collection instruments. This minimised confirmatory bias, and the audit trail showing how the items were improved using qualitative comment lent rigor to the

process of item improvement. This strengthened the validity of the final ‘excellent’ rating given to the Tool.

The inclusion of women is also rare in the content validity stage. This study fulfils both policy and research agendas that currently emphasise the need to ensure that women’s voices are heard at every stage of the research process. Their contribution to the development of items to measure the involvement of women was invaluable.

There are a number of ways this Tool can be used in practice and research. In its current from it could be used assess the implementation of evidence and target interventions to support skills development. As a research Tool it could be used to gather data on care processes to normalise birth. It could support efforts to establish relationships between the use of approaches associated with reduced interventions and outcomes.

Reference List

American Educational Research Association, American Psychological Association and National Council on Measurement in Education. (1999) ' Standards for educational and psychological testing ', In Streiner, D.L., Norman, G.R and Cairney, J (2015) *Health Measurement Scales: a practical guide to their development and use*. 5thedn. New York: Oxford University Press, pp.350-352.

Association of Internet Researchers (2012) *Ethical decision-making and Internet Research: Recommendations from the AoIR Ethics Working Committee (Version 2.0)* Available at http://www.aoir.org/reports/ethics2.pdf: (Accessed on 20/02/2015).

Betrán, A.P, Ye, J, Moller, A.B, Zhang, J, Gülmezoglu, A.M, Torloni, M.R. (2016) 'The Increasing Trend in Caesarean-Section Rates: Global, Regional and National Estimates: 1990-2014'.PLoSONE11 (2):e0148343.doi:10.1371/journal.pone.0148343

Birth Choice UK (2015) Available at http://www.birthchoiceuk.com/(Accessed on 10/02/2015)

Brocklehurst, P., Hardy, P., Hollowell, J., Linsell, L., Macfarlane, A., McCourt, C., Marlow, N., Miller, A., Newburn, M., Petrou, S., Puddicombe, D., Redshaw, M., Rowe, R., Sandall, J., Silverton, L., Stewart, M. and Birthplace in England Collaborative Group (2012; 2011) 'Perinatal and maternal outcomes by planned place of birth for healthy women with low risk pregnancies: the Birthplace in England national prospective cohort study', *BMJ: British Medical Journal*, 344 (7840), pp.17-17.

Coxon K, Sandall J and Fulop, N.J. (2014) 'To what extent are women free to choose where to give birth? How discourses of risk, blame and responsibility influence birth place decisions', *Health, Risk and Society*, 16(1), pp 51-67.doi.org/10.1080/13698575.2013.859231

Cicchetti, D.V and Sparrow, S (1981) 'Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behaviour', American Journal of Mental Deficiency, 86, pp.127-137.

Data protection Act 1998(United Kingdom). Available at http://www.legislation.gov.uk/ukpga/1998/29/contents (Accessed on 24/01/2015)

Darling, F (2015) ' Practitioners' views and barriers to implementation of the Keeping Birth Normal tool: A pilot study', British Journal of Midwifery 24 (7), pp 2-13

Economic and Social Research Council (2015) *Framework for Research Ethics.* Available at http://www.esrc.ac.uk/about-esrc/information/framework-for-research-ethics/index.aspx: (Accessed on 02/02/2015).

Fleiss, J.L. (1971) 'Measuring nominal scale agreements amongst many raters', Psychological Bulletin, 76(5), pp.378-382.

Grant J.S. and Davis, L. L. (1997) 'Selection and use of content experts for instrument development', *Research in Nursing and Health*', 20, pp. 269-274.

Green, J.M. (2012) 'Integrating women's view into maternity care research and practice', *Birth: Issues in Perinatal Care,* 39 (4), pp. 291-295.

Green, J. and Thorogood, N. (2004) 'Analysing qualitative data' in Green, J. and Thorogood, N. 1st edn. *Qualitative Methods for Health Research*. Great Britain: Sage, pp.173-200.

Greenhalgh, T. (2014) 'Evidence-based medicine: a movement in crisis', *British Medical Journal*, pp.1-7.doi: 10.1136/bmj.g3725.

Haynes, S. N., Richard, D. C. S. and Kubany, E. S. (1995) 'Content Validity in Psychological Assessment: A functional Approach in Concepts and Methods', *Psychological Assessment*, 7 (3), pp. 238-247.

Hatem, M, Sandall, J.Devane, D.Soltani, H. and Gates, S. (2009) 'Midwife-Led Versus Other Models of Care for Childbearing Women ' *Cochrane Database of Systematic Reviews*, Issue 4.Art No.CD004667, 10.1002/14651858.pub2.

Health Research Authority (2015) *UK Policy Framework for Health and Social Care Research.* Available at http://www.hra.nhs.uk/about-the-hra/our-plans-and-projects/replacing-research-governance-framework/ (Accessed on 04/04/2015).

Hospital Episode Statistics. (2012) Available at <http: // www.hscic.gov.uk/hes. (Accessed on 10/01/2015)

Hollowell et al (2015) 'The Birthplace in England Prospective Cohort Study: further analyses to enhance policy and service delivery decision-making for planned place of birth'. *Health Service and Delivery Research*. 3(36), doi 10.3310/hsdr03360

Involve (2012) *Briefing note six: Who should I involve and how do I find people?* Available at http:// www.invo.org.uk/posttyperesource/how-to-find-people-toinvolve (Accessed on 05/04/2015).

Kane, M.T. (2013a) 'Validating the Interpretations and Uses of Test Scores', *Journal of Educational Measurement,* 50 (1), pp. 1-73.

Kennedy, H. P., Grant, J., Walton, C., Shaw-Battista, J. and Sandall, J. (2010) 'Normalizing Birth in England: A Qualitative Study', *Journal of Midwifery & Women's Health,* 55 (3), pp. 262-269.

Kings Fund (2008) *Safe Birth: An independent inquiry into the safety of maternity services in England.* Available at http://www.kingsfund.org.uk/publications/improving-safety-maternity-services (Accessed on 10/04/2015).

Kings Fund (2015) *Better Value for the National Health Service*. Available at http:// www.kingsfund.org.uk/.../better-value-nhs-Kings-Fund-July%202015.pdf (Accessed on 02/02/2015).

Lynn, M. R. (1985) 'Determination and Quantification of Content Validity', *Nurs Res*, 35 (6), pp. 382.

McCourt, C., Rayment, J., Rance, S. and Sandall, J. (2012) 'Organisational strategies and midwives' readiness to provide care for out of hospital births: An analysis from the Birthplace organisational case studies', *Midwifery,* 28 (5), pp. 636-645.

Messick, S. (1989) 'Meaning and Values in Test Validation: The Science and Ethics of Assessment, *Educational Researcher*', 18, pp.5-11. Doi: 10.3102/0013189X018002005

Morgan, D.L. (2014) 'Pragmatism as a paradigm for social research ', *Qualitative Inquiry,* 20 (8), pp. 1045-1053.

Morse, J. M., Barrett, M., Mayan, M., Olson, K. and Spiers, J. (2008) 'Verification strategies for establishing reliability and validity in qualitative research', *International Journal of Qualitative Methods,* 1 (2), pp. 13-22.

National Institute of Clinical Excellence. (2007) *Intrapartum Care: care of the healthy women and babies in childbirth*. Available at: http: //www.nice.org.uk/guidance/cg55. (Accessed on 04/12/2014).

National Institute of Clinical Excellence. (2014) *Intrapartum Care: care of the healthy women and babies in childbirth*. Available at: http: //www.nice.org.uk/guidance/cg55. (Accessed on 04/12/2014).

O'Connell, R. and Downe, S. (2009) 'A metasynthesis of midwives' experience of hospital practice in publicly funded settings: compliance, resistance and authenticity', *Health,* 13 (6), pp. 589-609.

Polit, D.F. and Beck, C.T. (2006) 'The Content Validity Index: Are you sure you know what's being reported? Critique and Recommendations', *Research in Nursing and Health,* 29 pp. 489-497.

Polit, D.F., Beck, C.T and Owen, S.V. (2007) 'Is the CVI an acceptable indicator of content validity? Appraisal and Recommendations', *Research in Nursing and Health,* 30, pp.459-467.

Ramón E., White J., Beeckman K., Firth, L., Leon-Laris, F., Loytyed, C., Lubyen. A., Sinclair, M., Teijlingen (2015) ' Assessing the performance of maternity care in Europe: a critical exploration of tools and indicators) ', *BMC Health Services Research*, 15:491, pp 1-13
Doi: 10.1186/s12913-015-1151-2

Royal College of Obstetricians and Gynecologist (2015) Patterns of Maternity Care in English NHS Hospitals. Available at https://www.rcog.org.uk/globalassets/documents/guidelines/research--audit/maternity-indicators-2013-14_report2.pdf (Accessed on 02/04/2015)

Sireci, S. and Faulkner-Bond, M. (2014) 'Validity evidence based on test content', *Psicothema,* 26 (1), pp. 100-107.

Stemler, S. (2004) A comparison of consensus, consistency and measurement approaches to estimating interrater reliability. Available at http://www.pareonline.net/getvn.asp?v=9&n=4 (Accessed on 10/05/2015)

Stones, W. and Arulkumaran, S. 'Health-care professionals in midwifery care', *The Lancet*, 384 (9949), pp. 1169-1170.

Streiner, D.L., Norman, G.R and Cairney, J. (2015) 'Bias in Responding', in Streiner, D.L., Norman, G.R and Cairney, J. *Health Measurement Scales: a practical guide to their development and use*. 5thedn. New York: Oxford University Press, pp.100-125.

Teijlingen ER, Hundley V and Rennie A-M (2003) 'Maternity satisfaction studies and their limitations: 'What is, must still be best', Birth 2003, 30 (2), pp. 75–82.

Teijliingen, E. R. and Hundley, V. (2001) The importance of pilot studies. Available at http://www.sru.soc.surrey.ac.uk/SRU35.html. (Accessed on 20/05/2015).

Ye, J., Betrán, A., Pilar, Guerrero Vela, M., Souza, J., Paulo and Zhang, J. (2014) 'Searching for the Optimal Rate of Medically Necessary Cesarian Delivery', *Birth: Issues in Perinatal Care*, 41 (3), pp. 237-244.

Walsh, D. and Devane, D. (2012) 'A Metasynthesis of Midwife-Led Care', *Qualitative Health Research*, 22 (7), pp. 897-910.

World Health Organisation (2010) *The Global Numbers and Cost of Additionally Needed and Unnecessary Caesarian Sections Performed per Year: Overuse as a Barrier to Universal Coverage.* Available at http://www.who.int/healthsystems/topics/financing/healthreport/30C-sectioncosts.pdf (Accessed on 12/12/2014).

Figures and tables

Figure 1: Kane's Interpretation and Use Framework

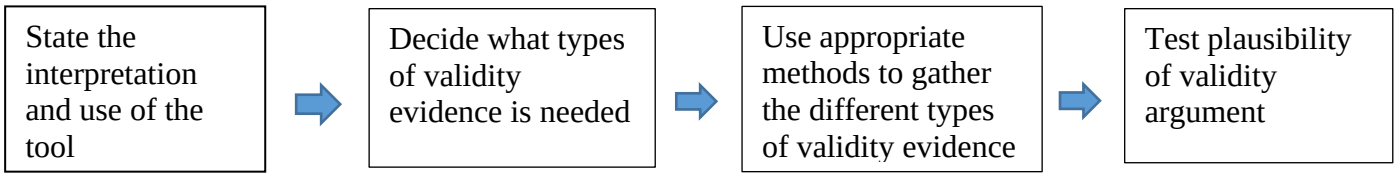

Table 2: A Description of Content Experts

Experts		**Women**	
EP1	Midwife, research in normal birth locally and internationally working with the WHO and Europe, Academic. Widely published/Author	**EW1**	MSLC, AIMS, Revaluing Care network
EP2	Midwife, research into midwife –led care. Academic role on Knowledge Transfer. Widely published/Author	**EW2**	MSLC Chair, AIMS and Association of Breast-feeding Mothers
EP3	Midwife, research in normal birth, Chair of Nursing and Midwifery, Widely published.	**EW3**	MSLC member, Doula
EP4	Midwife, Research Interest in patient-centred outcome, Organisational Systems and Cultures, Qualitative Methodology. Published.	**EW4**	MSLC Chair
EP5	Works with the Royal College of Midwives to promote midwife-led care. Widely published - 260 indexed publications and author/editor of 24 books.	**EW5**	MSLC member, Leads a birth choices group
EP6	Midwife, Researcher-PhD thesis on developing a tool, Education Project Manager. Widely published.	**EW6**	MSLC Chair
EP7	Midwife, lecturer, research interest-Art of Midwifery and Spirituality. Published.	**EW7**	Doula, Author

		EW8	MSLC Chair, Healthwatch Representative

Table 3: Evaluation criteria using the modified Kappa (k*)

I-CVI-adjusted for chance agreements (k*)	Criteria for evaluation
O.40-0.59	Fair
0.60-0.74	Good
>0.74	Excellent
>0.78	Excellent - Cut-off proposed by Polit et al., 2007, and used in this study

Table 4: shows the I-CVI for each item

Item	I-CVI	Item	I-CVI	Item	I-CVI	Item	I-CVI	Item	I-CVI	Item	1-CVI
1.1	0.93	4.4	1.0	7.1	1.0	8.8	0.78	9.9	0.85	12.1	1.0
1.2	0.93	5.1	0.40	7.2	1.0	9.1	1.0	10.1	0.78	12.2	1.0
2.1	0.93	5.2	0.38	8.1	0.93	9.2	0.93	10.2	0.78	12.3	1.0
2.2	1.0	5.3	1.0	8.2	0.78	9.3	0.85	10.3	0.86	12.4	1.0
2.3	0.93	5.4	0.46	8.3	0.46	9.4	0.93	10.4	1.0	12.5	1.0
3.1	0.86	6.1	0.93	8.4	0.64	9.5	1.0	11.1	1.0		
4.1	0.93	6.2	0.93	8.5	0.86	9.6	1.0	11.2	1.0		
4.2	0.86	6.3	1.0	8.6	0.93	9.7	0.84	11.3	0.93		
4.3	0.93	6.4	0.85	8.7	0.64	9.8	0.64	11.4	1.0		

Pink -poor; Blue-fair; Green-Moderate; White-excellent

Table 5: Items for deletion and improvement based on I-CVI

Item	I-CVI adjusted for chance agreements	Evaluation based on standards developed by Polit et al. (2007) using Fleiss (1981) and Cicchetti and Sparrow (1981) as a guide	Decision
5.1	0.40	Poor	Item deleted
5.2	0.38	Poor	Item deleted
5.4	0.46	Fair	Item Improved
8.3	0.46	Fair	Item Improved
8.4	0.64	Good	Item Improved
8.7	0.64	Good	Item Improved
9.8	0.64	Good	Item Improved

Table 6: Themes derived from Qualitative Comments

Categories	**Themes**
1. Evidenced-informed-Therapeutic Alliance	Theme 1: Inclusion of women in decision-making
2. Evidenced-informed-Reducing Unnecessary Interventions	Theme 2: Surveillance increases risk of unnecessary interventions
	Theme 2a: Quality of evidence 2.1a Evidence to support normal/physiological birth 2.2a. Evidence to support women in labour 2.3a. New evidence
3. Language	Theme 3a: Medicalised language Theme 3b: Clarity