TASMANIAN SCHOOL OF BUSINESS & ECONOMICS

October 2014

Improving Inter-Assessor Reliability for Health Service Accreditation:

A Literature Review

Australian Institute of Health Service Management

AUSTRALIAN COMMISSION ON SAFETY AND QUALITY IN HEALTH CARE

Edward Benecke (ID: 207-290)
ebenecke@postoffice.utas.edu.au

Suggested citation:
Benecke, E. (2014). Improving Inter-Assessor Reliability for Health Service Accreditation: A Literature Review, University of Tasmania, School of Business & Economics

Abstract

Background

Health service accreditation programs are designed to assess performance against predetermined standards and used as a driver of safety and quality in health care. Despite significant investment in accreditation programs inter-assessor reliability is difficult to achieve and is a key factor affecting accreditation outcomes. Low reliability is seen to disengage health professionals and managers and a lack of certainty about what is being assessed compromises the validity of the whole accreditation process.

Aim

Conduct a literature review identifying best practice strategies to improve the inter-assessor reliability of surveyors during accreditation to ensure the consistent assessment of this application of standards and to promote confidence in the Australian Health Service Safety and Quality Accreditation (AHSSQA) Scheme.

Methodology

Relevant key terms were search in selected electronic databases. Results were assessed against exclusion criteria. This review covers the literature from multiple disciplines published since 2007 to prevent duplicating a prior review conducted by Greenfield, Pawsey and Braithwaite (2007). The content of primary search results was analysed to refine studies to those sufficiently relevant to inter-assessor reliability.

Findings

Five broad themes relating to improving inter-assessor reliability were identified within the literature and studies were grouped by theme for summary and analysis. The key themes are: workforce management and surveyor selection; assessor experience; training and continued professional development; assessor teams; and regular auditing.

Conclusions/Implications

The review determined that inter-assessor reliability could be promoted through sufficient assessor workforce management, selecting proficient experienced assessors, providing appropriate training and education in assessment criteria, promoting team surveying and regular auditing. The application of best practice strategies will ensure consistent application of NSQHS Standards & promote confidence in the Australian Health Service Safety and Quality Accreditation (AHSSQA) Scheme. Further research, specifically in health care accreditation, is required to test and develop these strategies.

Contents

1.0 Introduction

Accreditation is a formal declaration by a designated authority that an organisation has met predetermined standards (Greenfield, Pawsey & Braithwaite 2011). International health service accreditation is widely used as a driver of safety and quality. Accreditation programs are designed to monitor and assess performance against predetermined standards. The standards are implemented adaptively and dynamically in response to varied contextual stimuli arising from different health system domains. (Hinchcliff et al. 2012; Hinchcliff et al. 2013).

Assessors have a central role in health care accreditation programs. Accreditation assessors are typically a unique set of individuals who come together for a short period of time to complete the task of assessing a health organisation (Greenfield et al. 2009). The assessor "role involves educated and trained agents considering the merits and demerits of the enterprise being subject to judgement" (Greenfield, Braithwaite & Pawsey 2008 p435).

Despite significant investment in health care accreditation and programs as important drivers to improve quality and safety in health care, further research and a transparent examination of different aspects of accreditation is required (Greenfield, Braithwaite & Pawsey 2008; Greenfield & Braithwaite 2009).

A key factor affecting accreditation outcomes is inter-assessor reliability. Assessor findings can be said to be reliable if the process and assessment is consistent between individual assessors (Greenfield, Braithwaite & Pawsey 2008). Inter-assessor reliability is typically difficult to achieve in assessment programs (Jayasinghe, Marsh & Bond 2006), hence there is a critical challenge to ensure assessor judgements are consistent (Greenfield et al. 2009).

The inappropriate or inconsistent surveying by assessors has the capacity to undermine health services' confidence in accreditation programs. Accreditation agencies claim that assessor reliability is achieved through defined assessor selection criteria and training programs (Greenfield et al. 2009, pg 113). However, studies have identified inconsistent assessor behaviour, competencies and conduct which have compromised the reliability of surveys. Siggins Miller (2009) recognised that for assessors to maintain competency a minimum number of surveys and training hours need to be completed each year.

Low level inter-assessor reliability within accreditation programs is seen to disengage frontline health professionals and managers (Hinchcliff et al. 2013). A lack of certainty about what is being assessed compromises the validity of the whole accreditation process.

Aims and objectives of this review:

The aim of this literature review is to propose best practice strategies based on interpretation of the results of studies included in the review that can be incorporated into health service accreditation programs to improve inter-assessor reliability.

2.0 Health care Accreditation in Australia

2.1 The National Safety and Quality Health Service Standards

The Australian Commission on Safety and Quality in Health Care (ACSQHC) developed the ten National Safety and Quality Health Service (NSQHS) Standards (Appendix 1) following consultation and collaboration with State and Territory jurisdictions, technical experts and a wide range of stakeholders including health professionals and patients (ACSQHC 2011). The national program is recognised as an important driver in safety and quality improvement, achieved by applying standards and promoting the uptake of evidence base clinical and organisational practices (Hinchliff et al. 2013).

The primary aims of the NSQHS Standards are to protect patients and to improve the quality of health service provision. Health service organisations can also use the Standards as an improvement mechanism to recognise developmental goals (ACSQHC 2011). The Standards were designed for use by all health services and to provide assurance that minimum standards of safety and quality are met. Compliance to the standards through accreditation processes has commenced in hospitals, day procedure centres and dental services throughout Australia.

Prior to the NSQHS Standards there were multiple accreditation bodies assessing health services against safety and quality standards in Australia. The development of the NSQHS Standards and the endorsing of a national accreditation scheme, coordinated by the ACSQHC, by the Australian Health Ministers, ensured consistent assessment by each accrediting agency (Adams 2011).

2.2 The Australia Health Service Safety and Quality Accreditation Scheme

The Australian Health Service Safety and Quality Accreditation (AHSSQA) Scheme was launched in January 2013. "The Scheme outlines the role of key groups operating the new Australian accreditation system. They include:

- Health Ministers, who endorse the standards.
- The ACSQHC who approve accrediting agencies and review the safety and quality implications that flow from accreditation.
- Regulators that mandate standards oversee accreditation program content and receive accreditation data.
- Health service organisations that continue to select the accrediting agency to assess their organisation and meet the NSQHS Standards.
- Accrediting agencies that assess services and provide information from accreditation to regulators and the Commission." (Adams 2011 p.3)

The credibility and reliability of accreditation programs depends on the management of assessors by an accrediting organisation (Bohigas et al 1998), as such accrediting agencies that wish to accredit health service organisations to the NSQHS Standards must undertake an application and assessment process conducted by the ACSQHC (ACSQHC 2012). A list of currently approved accrediting agencies is at Appendix 2.
A requirement of the approval to become an accrediting agency is to enhance surveyor training and performance management to increase the reliability and validity of

assessment processes, including participation in surveyor training convened by the ACSQHC from time to time (ACSQHC 2012).

To be approved, accrediting agencies need also to be accredited by an internationally recognised body such as the International Society for Quality in Health Care (ISQua) or the Joint Accreditation Scheme of Australia and New Zealand (JAS-ANZ). In addition to providing accreditation to accreditation agencies, JAS-ANZ and ISQua also provide guidelines and recommendations for the selection of assessors with appropriate expertise and experience; relevant training and support for assessors; and encouragement of continuing professional development, and the assessment of assessors (JAS-ANZ 2012; ISQua 2012; Siggins Miller 2009).

3.0 Inter-Assessor Reliability

Health care assessors are typically health professionals who are trained and skilled in surveying techniques, and gather the relevant information to enable a hospital's (health service) compliance with a set of standards to be assessed (Bohigas et al 1998). In health care it is critical for reliable judgement to be made as assessors are linchpins in accreditation schemes (Hurst 1997).

Several reasons have been cited and promoted as capable of undermining inter-assessor reliability. These include:

- the introduction of opinions combined with interpretation of valid and relevant external evidence (Greenfield et al. 2013);
- sporadic exposure to the accreditation process, the cost and frequency of training required to maintain skill, the challenges of a part-time workforce and the difficulty of measuring reliability (Siggins Miller 2009);
- a lack of transparent interpretation processes for the standards and assessor workforces without appropriate capacity (Hinchcliff et al. 2013);
- accreditation personnel who did not have a strong understanding of the health sector, accreditation processes and negotiation and report writings skills (Greenfield et al. 2009);
- assessors reviewing only a few health services give less reliable ratings than those who review a large number (Jayasinghe, Marsh, Bond 2006);
- assessor preferences weakening the accreditation process by introducing bias (Hurst 1997).

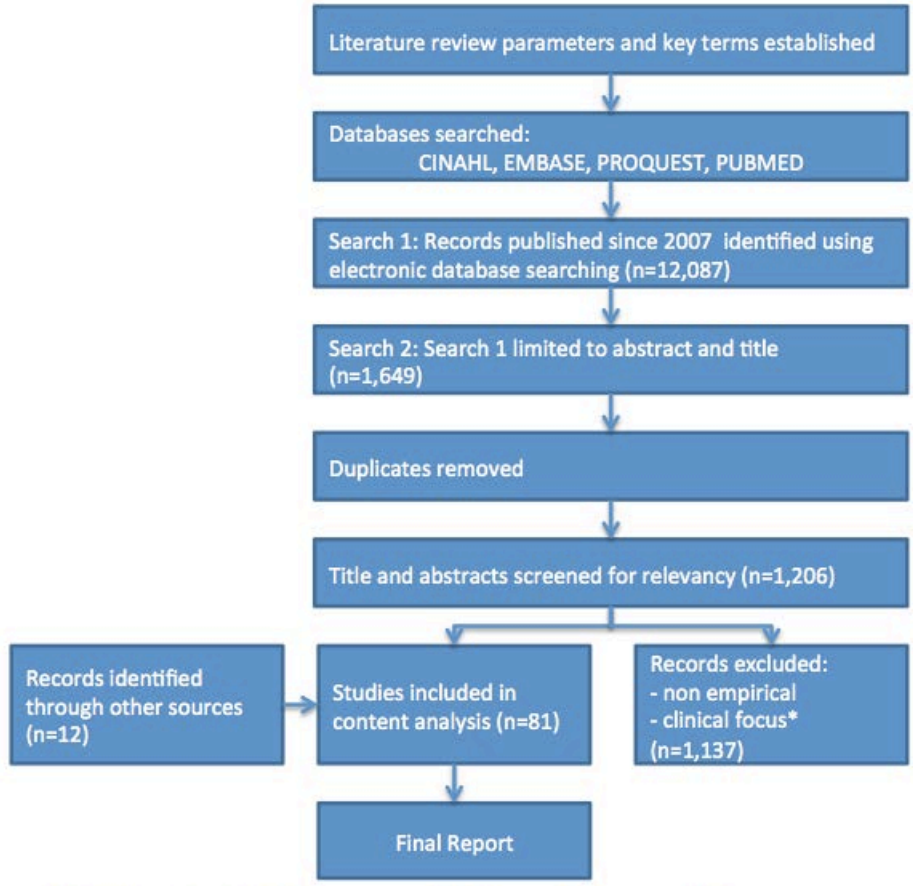
Consistency is difficult to achieve and this is a concern when individuals have to make accreditation assessments (Greenfield et al. 2013). Therefore, identifying best practice to improve inter-assessor reliability is required for the policy decision makers and researchers to develop and evaluate accreditation programs.

4.0 Methodology

4.1 Literature Aggregation

A multi-step search strategy was undertaken to compile relevant literature for the review (see Figure 1). Four major multi-discipline electronic databases were interrogated in September 2014. Cumulative Index to Nursing and Allied Health Literature (CINAHL), PubMed, Proquest Central and Embase were chosen to ensure literature from a variety of disciplines were included for exploration.

Figure 1: Flowchart of search strategy and relevance assessment



** Clinical studies excluded if inter-rater reliability used only as a validation measure.*

Trial searches were used to develop the three key word groupings in Table 1, which were searched concurrently in each database. The use of numerous synonyms for inter-assessor, reliability and accreditation was essential for maximum recall, producing a larger number and wider range of references relevant to the review. Boolean phrases and truncated search terms were utilised. The substitutes for inter-assessor were searched as one and two words to capture maximum results.

The combined database searching yielded 24,127 (see Table 2). Greenfield, Pawsey and Braithwaite published in 2007 *'Intra-rater and inter-rater reliability in health care accreditation: Literature review'* which systematically identified and analysed publications concerned with intra-rater and inter-rater reliability. To prevent duplication

of this earlier work this literature review will review the literature on inter-rater reliability since Greenfield, Pawsey and Braithwaite's earlier review. Therefore a further filter was applied to the results within each database removing references published prior to 2007. The results were then restricted to find only articles with the search terms present in the title and abstract to enhance precision and capture more highly relevant articles.

Table 1: Database Search Terms

Inter-Assessor		Reliability	Accreditation
inter-assessor	inter-examiner	reliability	accredit*
interassessor	interexaminer	variability	certif*
"inter assessor"	"inter examiner"	variation	audit*
inter-rater	inter-observer	consisten*	
interrater	interobserver	agreement	
"inter rater"	"inter observer"	validity	
inter-surveyor	inter-evaluator		
intersurveyor	interevaluator		
"inter surveyor"	"inter evaluator"		

**Truncated search terms.*

The ensuing 1,649 references were downloaded to reference management software (Endnote X7.1) to combine the results and remove duplicates. The named author screened titles and abstracts of the 1,206 records manually to identify the articles to include for content analysis. This method was selected because test searches using a fourth group of keywords to narrow the results to studies which focused on improvement or enhancement of inter-assessor reliability demonstrated that this resulted in exclusion of relevant studies.

Table 2: Database Results: Number of Articles

Search Parameters	CINAHL	EMBASE	PROQUEST	PUBMED	TOTAL
Key Terms	558	925	21,387	1,257	**24,127**
Results since 2007	286	615	10,497	689	**12,087**
Abstract & Title	287	525	224	613	**1,649**
Duplicates*					**443**
				References screened	**1,206**

**Duplicate records were removed using Endnote X7.1 reference management software.*

The exclusion criteria used to screen the netted search results were: non-English language, non-empirical studies, unrelated clinical focus or articles that merely measured or reported on inter-assessor reliability to validate a specific practice or therapy.

In addition to the database searching, other publications focusing on the evaluation of inter-surveyor reliability in health care accreditation that were not captured in the database results were included in the review. Known key authors were searched in Google Scholar to find relevant articles since 2007. The extra literature was primarily

from three supplementary journals not contained in the major databases: BMJ Safety and Quality, Health Information Management Journal, and the International Journal for Quality in Health Care.

4.2 Content Analysis

Eighty-one studies were included in the content analysis and were assessed in detail. Upon further reading of the texts, where the reference to inter-assessor reliability was not sufficiently relevant to the aim of this review, the study was not included in the thematic analysis.

By focusing on connections amongst the empirical findings of the final 38 studies a number of themes were identified. The themes helped to provide a bridge between the scatter and assortment of research on the topic and integrate the results of multiple studies (Baumeister & Leary 1997). The thematic analysis was key to organising the evidence and shaping the strategies proposed to improve inter-assessor reliability. Methodologically diverse studies meant that meta-analysis (aggregation of results) was not possible.

To prevent a loss in value of the thematic analysis, sufficient detail is presented for each study for readers who may be sceptical of the conclusions or who prefer to make their own determinations. The reader is able to evaluate the strength of the method, results and context of the study rather than simply relying on the named author's interpretation of the conclusion (Baumeister & Leary 1997, Thomas & Harden 2008).

5.0 Thematic Analysis

5.1 Previous Research

Greenfield (et al. 2007) completed a comprehensive search and multi-analysis of the literature on inter-rater and intra-rater reliability in health care accreditation in August 2007. Their analysis noted that the reliability of accreditation surveyors remains an under researched issued and that research into reliability is concentrated on statistical measurement of inter-rater reliability in clinical care. Only two studies were identified that argued for the improvement of inter-rater reliability through initial and ongoing training of surveyors. Greenfield (et al. 2007) concluded that empirical studies of inter-rater reliability of accreditation surveyors remain to be done and further research is necessary.

5.2 Key Themes

5.1.1 Workforce Management

The importance of workforce management was identified from the literature, with studies highlighting the importance of assessor selection, and the requirements for appropriate leadership and rater competency.

Greenfield (et al. 2009) collected and thematically analysed data from 193 health care accreditation stakeholders using focus groups, interviews and a questionnaire. Two

factors identified by the participants, namely accreditation agency personnel and the renewal of the surveyor workforce, emphasised the importance of accreditation survey workforce management to promote inter-assessor reliability. Participants suggested accrediting agency personnel required an understanding of the accreditation process, health sector and strong facilitation, relationship management and report writing skills. It was believed that employees with a defined skillset would promote reliability of the survey. Participants also identified the selection process for health professionals to train as surveyors as a dynamic that shaped survey reliability. Selecting the appropriate surveyors aims to ensure they have explicit and tacit knowledge of the health organisations they assess. The importance of managing the accreditation workforce was highlighted by participants remarking that surveys performed in quick succession can result in details from different facilities being confused and without adequate time between reviews surveyors can be under or unprepared for surveys. All of these factors may undermine reliability. Other concerns were raised about balancing the demands for a surveyor workforce and the need to mitigate potentially overworking a defined workforce.

A trial of two differing accreditation surveys was conducted congruently in 17 acute health care organisations (Greenfield et al. 2012). Feedback was collected from the organisations staff, surveyors and patients identifying the need for appropriate workforce management of surveyors: the feedback from surveyors was that inter-assessor reliability and consistency of surveys would have improved with the provision of clearer information and the allowance of sufficient preparation time.

A study by Hinchcliff (et al. 2013) collected and analysed data from focus groups and interviews of 258 health care stakeholders across Australia regarding perspectives on implementing accreditation programs. The stakeholders selected were said to have significant knowledge of the accreditation process. The study identified the level of inter-surveyor reliability as a key barrier for implementing a valid health service accreditation program. Suggested factors to improve inter-rater reliability were transparency of processes of external auditing organisations and the need for workforces with appropriate capacity. The study also identified the need for health care leaders to champion quality and safety improvements and foster engagement in assessments, rather than simple pragmatic surveying. Distributed leadership would also help to share experiences.

Although the studies by Greenfield (et al. 2009; et al. 2012) and Hinchcliff (et al. 2013) used comprehensive methods to collect and analyse data from a range of key stakeholders, the outcomes remain based on the perceptions of participants rather than actual data to test the perceptions.

Kott (et al. 2008) illustrated the use of statistical measurement of inter-rater reliability as a workforce management and assessor selection method. The study statistically assessed a rater selection methodology to improve ratings precision and rater quality. In attempting to validate assessment and improve inter-rater reliability 284 raters were drafted for the study based on the minimum education and experience thresholds. Comprehensive training was provided to the raters who then proceeded to make mock assessments. The assessment ratings were evaluated for competency using percentage agreement of rater scores with an expert panel consensus and inter-rater reliability calculated using the Cicchetti kappa method. The kappa values revealed a significant difference among rater

groups. Rater selection was based on an 80% concordance with the expert panel consensus for competency and an >0.85 kappa value for inter-rater reliability. A Cochrane Q test was used to calculate a composite score for raters that were both competent and reliable. The result was that only 223 raters were selected from the original pool. The introduction of similar statistical parameters for workforce selection by accrediting agencies may improve the workforce competency and inter-rater reliability.

5.1.2 Assessor Experience & Personality

In addition to an increased emphasis on workforce management and selection to improve inter-assessor reliability, the literature review included studies that evaluated the importance and impact of experience on inter-assessor reliability whilst conducting assessments.

Kulkarni, Walker & Carter (2014) assessed the inter-rater reliability in traumatic brain injury classifications (TBI) using retrospective case review. Four raters were compared pair-wise for reliability using the kappa statistical method and percent concordance. Rater concordance was higher in raters with more years practice and experience in TBI compared with less experienced colleagues. This study highlights that raters with more experience may be more likely to correctly and reliably score assessments, however only four raters were assessed which limits the weight of the study.

Olsen (et al. 2014) examined inter-rater reliability during physical examinations of horses. The study evaluated the subjective assessment of neurological deficits in horses by calculating the reliability of data. The raters included board-certified surgeons, second year residents and interns. A median of five raters assessed 25 horses, completing a questionnaire for each step of the examination. Inter-rater agreement was quantified using the intraclass correlation coefficient. The results showed that the more experienced raters were more consistent in their assessment scoring than residents, interns and students. Despite the fact that only a small number of raters were assessed, the results support the proposition that more experienced assessors have greater inter-assessor reliability.

The interobserver variability between radiologists evaluating complex computed coronary angiographs with varying levels of experience was assessed (Kerl et al. 2012). The differences between two board-certified radiologists with eight and ten years experience and two radiology residents with one and three years experience were evaluated. The correlation coefficient was positive and matched for raters with 8 and 10 years experience. Where as, the results for the two less experienced residents was significantly lower, demonstrating that level of experience influences the observer variation. The study is limited as it only assessed four study participants, therefore the results may reflect on the individual's competencies as opposed to being a sample reflecting population trends.

Miller (et al. 2011) evaluated personality traits of raters relative to their psychopathy scoring tendencies. Twenty-two graduate, masters and doctoral level students from two American psychology programs independently scored four criminal offenders following a training session on a clinician-scored instrument for the assessment of psychopathy. The results showed a significant variability between the raters conducting the assessment. The results were cross-referenced with raters' self reported personalities. Raters' personalities

offer insight into some variability of the scores assigned to offenders. For instance, compassionate more agreeable raters tendered to rate offenders as less psychopathic. Overall the results supported the conclusion that inter-rater reliability and scoring tendencies can be influenced by personality traits. The study had some limitations including assessment of only four cases and that the majority of raters were graduate students, however it highlights the need to mitigate the opportunity for personalities to influence objective evaluations. Health care assessors who impose personal values may pose a significant threat and undermine the validity of the accreditation process.

Conversely to the aforementioned studies a study by Beckman (et al. 2014) reviewed the inter-rater reliability of a specific obstetric triage system and demonstrated that experience had only a limited effect on assessment. The audit assessed the inter-rater reliability of thirty random midwives with varying degrees of triage experience. The participants underwent training and assessed fictional scenarios. The results, calculated using Cohen's kappa, showed a reliability range from 68% - 100%. Evaluating the assessor demographics showed that for trained raters there was no significant correlation between rater experience and the variation in assessment.

Another study retrospectively analysed the relationship between experience level and inter-rater reliability and reported little effect (Kott & Swartz 2012). The inter-rater reliability of 284 and 367 raters from two earlier studies investigating the use of the Unified Parkinson's Disease Rating Scale was examined. The combined 651 raters received training on the scale as part of the previous studies. No significant correlations were identified between kappa results and the number of years of clinical experience. These results suggest that rater experience may be less influential on inter-rater reliability in assessments when training has occurred. As such, experience levels may not be an appropriate means to grade and select raters.

A large single centre study examined the variability in interpretation of electroencephalography (EEG) results by six board-certified clinical neurophysiologists at one institution (Chari et al. 2012). Measured by Cohens kappa coefficient inter-rater reliability ranged from 0.29 to 0.64. The wide-ranging results demonstrated the subjectivity of interpretations even among highly qualified and experienced clinicians at one institution. The researchers suggest that the development of consensus guidelines may improve inter-rater reliability with the review of controversial ratings amongst peer groups in order to arrive at a consensus.

Furthermore, Smith (et al. 2013) concluded that raters with a variety of background experience could achieve a high level of inter-rater reliability. The study assessed the reliability of raters using a functional movement screen. Raters with different experience and educational backgrounds including two physical therapy students with varying scale assessment experience, a certified rater and a postdoctoral researcher in Biomechanics and Movement Science. Raters underwent a two-hour training session. The inter-rater reliability for each rater was high, with intraclass correlation coefficients ranging from 0.87-0.99. The results of the study supported the concept that regardless of experience professionals who receive appropriate training can administer and score assessments in a multi-disciplinary team with high inter-rater reliability. However, the study is limited in its practicality as only four raters assessed.

The available literature predominantly highlights that experience can impact inter-rater reliability, and is potentially correlated to the level of training assessors receive and, as such, is an important factor in the section of assessors. Additionally, the use of consensus guidelines may prevent the use of subjective assessment of interpretations and minimise variation resulting from experience levels.

5.1.3 Assessor Training & Continued Professional Development

Education interventions and training of assessors to promote inter-assessor reliability was a common theme identified in literature searching. Greenfield (et al. 2009) investigated the reliability of health care accreditation surveys using focus groups, interviews and surveys of key stakeholders. The perceptions of the participants was that inter-assessor reliability is achieved through standardised training which emphasises a universal surveyor role, promotes consensus interpretation of the accreditation standards and analogous survey team conduct. Participants also asserted that missed training could mean a surveyor was not aware of current standards which in turn would decrease the reliability of the assessor.

A study by Greenfield, Braithwaite & Pawsey (2008) also researched the typology of health care accreditation surveyors. The two-phase research first observed a small survey team during an accreditation review and secondly had a panel of surveyors review the resultant typology. Three unique surveyor styles were identified from the study, namely the interrogator, the explorer and the discusser. Greenfield, Braithwaite & Pawsey (2008) postulated that the identification of surveyor styles has practical relevance for surveyor training and development. Surveyor reliability may be promoted by pairing new surveyors with a similar style mentor to promote their learning. The confirmation of the styles by the surveyor panels did note that surveyors can transverse styles during a survey.

The effectiveness of a single educational intervention to reduce inter-observer variability in rectal cancer targeted volume delineation and further supported the use of training to improve reliability (Doughton et al. 2012). Specialists and trainees contoured a rectal cancer both before and after an educational intervention. Twenty-four matched pairs improved between pre and post education intervention by an average of 9%, showing the education intervention to be effective in improving inter-observer reliability.

Garcia-Reyes (et al. 2014) investigated the effect of education on MRI interpretation accuracy for prostate cancer diagnosis. Five radiology fellows underwent a dedicated education program and proceeded to evaluate prostate MRIs retrospectively for lesion detection. Accuracy of the lesion detection increased significantly post education with anterior prostate detection increasing by 42.8%. The results support the proposition that substantial inter-observer variability in MRI interpretation can be reduced with education programs.

A large sample study further supports rater training as a tool to improve inter-rater reliability and also compared effectiveness between online and composite training modules (Gaur et al. 2010). 104 raters from eight countries underwent training for two rating scales used in clinical trials. The study evaluated change in inter-rater reliability of raters who participated in the website certification training for scale one and a combination of online and face to face training for scale two. Inter-rater reliability was calculated using the kappa statistic. The results found that raters participating in the

website certification for scale one were poor with limited reliability. Whereas, scale two training sessions with expert trainers improved inter-rater reliability demonstrating that face to face training may have greater effectiveness. Potential prior experience and previous education for scale two may have skewed the results and affected the comparability of the reliability measurement. A study evaluating reliability with the same rating scale using the two training methods could provide insight into this.

Quigg and Lado (2009) examined the inter-rater reliability of content reviews for continuing medical education (CME) presentations. Twelve neurologists with limited experience in CME rated a Power-point presentation for mandated and important attributes for CME. A kappa statistic was 0.115 showing inter-rater reliability was in the poor range. The researchers deemed the rater's lack of training in CME exacerbated the lack of inter-rater reliability. The medical specialists were highly experienced in their field but not in the details of CME. Also the role of the specialists rating the CME presentations was not clearly defined, with some assuming editorial roles and others having a more impressionistic, laissez-faire approach. These findings demonstrate the need for rater training in health care accreditation and the necessity for raters to be informed of the standards and their role in applying them.

A study assessing auditory ratings showed that consensus training only marginally increased inter-rater reliability (Iwarsson & Peterson 2012). Thirteen students with basic level education and without clinical experience undertook four fortnightly consensus training sessions to examine whether high inter-rater reliability could be achieved in the assessment of voice quality. A prerequisite for rater reliability was that raters were able to identify voice qualities and apply the correct terminology associated with them. Cronbach's alpha was used to measure inter-rater reliability. The results from pre-training (α=0.913) to post training (α=0.932) were only marginal and less than hypothesised. The authors proposed that the raters who had previously received basic voice education might have been familiar with the reference voice materials. The study did show that the consensus training increased the ability of the listeners to identify voice quality, however, further research was recommended with the use of randomly selected reference material and a less homogenous rater group.

Lee & Choi (2014) assessed the inter-observer reliability of grayscale ultrasonography evaluations between radiologists. Three radiologists surveyed 91 phantom images. The surveyors first scored images and then underwent two hours of interactive education and proceeded to score more images. The inter-observer reliability was calculated using a kappa statistic showing a slight and moderate reliability before any education and a substantial reliability after the education for all participants. This study strongly supports the use of interactive education to improve inter-assessor reliability. However, the study is limited by its sample size of three participants.

Brixely, Guse & Gorelick (2010) investigated the reliability of child passenger safety restraint observations by trained community observers compared to certified technicians and a gold standard. Nine observers were shown 75 photos and assessed restraints for the appropriateness of the harness, applicability and safety. A kappa statistic of 0.28 showed overall poor inter-assessor reliability. In contrast to the results of other studies, trained community observers had a trend toward poorer judgment compared with certified observers in determining the harness appropriateness.

A study by Bell, Aldinger & Richey (2011) concluded that inter-rater reliability may not be correctable through education. Forty-one physicians and coding specialists reviewed a demographic survey containing mock coded patient documents to determine which most accurately described the amount, complexity and appropriateness of the work documented. Overall the results for all respondents showed there was little agreement in the code assignments between either professional group. This lack of concordance suggests that consistency, particularly amongst coding specialists, may not be correctable by training and education. The study concludes that an ongoing significant variability between raters is more likely an indicator of failure of the instrument or process.

Wimer (2007) investigated the inter-rater reliability of athletic training site accreditors. The accreditors were assigned to judge and interpret education programs in accordance with published standards. The 93 accreditors had greater than 10 years certified athletic training experience and had fulfilled the training requirements. Statistical results of inter-rater reliability showed that there was poor reliability in the judgments made by the accreditors, who reacted differently to identical information provided in the scenarios and disagreed with certain standards. However, although the study was designed to mimic real-life situations, it was conducted online using written scenarios and respondents did not conduct face-to-face interviews or work in collaboration with other trained site visitors. Wimer (2007) proposes that a knowledge test of the current standards should be given to all site visitors, as well as appropriate continuing education for quality improvement and promoting consistency among surveyors. The results of this study demonstrate that training methods may not result in inter-rater reliability. Feedback provided by the study participants highlighted that continual procedural changes without accompanying training complicated the accreditation process and created frustration within the workforce.

Numerous results from studies in various disciplines support the proposition that training and education are a key strategy to promote inter-rater reliability. However, despite assertive evidence, the question must be raised whether a similar educational intervention in a health care accreditation context will produce a reduction in rater variability as evidenced across disciplines.

5.1.4 Assessor Teams

Yao, Foster and Alrich (2009) investigated the inter-rater reliability of a team scored teacher portfolio required for initial teacher certification. Similar to health care accreditation the review teams need to interpret and judge large amounts of information in a consistent way. Eight review teams received several rounds of team level training in scoring the portfolios. Inter-rater reliability was statistically assessed and showed a strong level of reliability for the team portfolio scores. The results supported a strategy where raters work together in teams of two or three and use discussion to arrive at a negotiated score to improve rater consistency.

Schildmeijer (et al. 2014) reviewed the reliability of teams using a well-developed and consistent method to detect adverse health care events. Five teams from five hospitals in Sweden retrospectively reviewed records to identify adverse events using the global trigger tool. The results, calculated using combined kappa statistics, showed inter-rater reliability differing between the teams, ranging from 0.32 to 0.60 in the identification of

adverse events. No education or instructions were provided to the teams about reaching a consensus prior to the study. The team composition also varied from team to team within the study and may have influenced the reliability. The researchers suggested that group training may have achieved a better agreement across the teams. The large differences between the teams contradicts proposal that rating by a team improves the reliability of assessment. However, as only five teams were used it can't be presumed that this study is representative of other teams.

A study by Hutchinson (et al. 2010) compared the agreement levels between health care professions conducting retrospective case note reviews using two methods. The review was conducted in teams composed of doctors, specialist nurses, non-clinical audit staff and clinically trained audit staff. 684 randomly selected case notes from nine acute hospitals in England were reviewed. Same staff-type groupings showed moderate to good inter-rater reliability across the two methods, whereas multi-discipline review teams reviewing the same records had a weak to moderate inter-rater reliability. Hence, the composition and type of staff in review teams must be considered.

Greenfield (et al. 2013) conducted a study assessing the reliability of two four member hospital survey teams accrediting a large Australian teaching hospital. The members from each team were matched for experience. Data was collected through interview, questionnaires and observations from hospital staff, accreditation personal, surveyors and research staff. The results showed high levels of agreement between the survey teams' ratings, promoting peer reviewer team assessments as reliable. However, Greenfield et al. had difficulties conducting the study and suspected that the results were compromised. For instance, surveyors opted to collaborate and clarify interpretation and concerns because they were aware they were being scrutinised by the observing researchers.

5.1.5 Regular Auditing

The importance of validating or auditing accreditation processes was a common theme in the literature results. The reliability of the accreditation process and consistent application of the NSQHS Standards will promote inter-assessor reliability.

For instance Greenfield (et al. 2011) conducted an empirical study to determine the effectiveness of assessment of a short notice survey at twenty health care organisations. Thirty surveyors were provided training on short notice surveys. The study matched ratings from previous advanced notice surveys. Statistical tests were performed to identify any issues with the new surveying technique. The trial overall demonstrated a high level of inter-assessor agreement.

Abdel Baki (et al. 2012) studied the diagnostic inter-rater reliability of EEG interpretations among professional neurophysiologists and found that the level of agreement varied significantly. Further investigation showed that the level of agreement between raters was dependant on the category being assessed. The study postulated that diagnostic categories with little agreement among raters should be a priority for continuing education. Auditing the inter-rater reliability of health care accreditation surveyors for each individual standard may identify areas of misinterpretation and hence identify areas to emphasise in training.

Carriker & Isaacs (2013) developed a process to evaluate and score inter-rater reliability of individual auditors of a hand hygiene compliance program. The hand hygiene auditors were reviewed and observations documented by a hand hygiene compliance team. In the first audit 181 observations were collected with 9 discrepancies identified. The audit identified discrepancies between auditor observations and allowed for additional education and training if necessary. The validation process was repeated with continued improvement demonstrated. While observations can be influenced by subjectivity, a validation process can assist in personnel remaining objective.

Carson, Fitch and Vachon (2009) reported on the audit results of a palliative care assessment tool designed to identify areas where care needs were not being met. Multidisciplinary teams, patients and families from the palliative care unit in a large urban Canadian teaching hospital audited the assessment. The results demonstrated variability in the application of the care assessment tool, as there was inconsistency between the ratings of the same situation on two occasions by members of the same team. To have confidence in the findings from an assessment tool, that tool must be applied in a reliable manner and in the setting where it is being used.

A study by Jelovsek, Kow and Diwadkar (2013) aimed to determine which tools exist to best assess directly the psychomotor skills in medical trainees. Thirty known tools available for the assessment of psychomotor skills in medical trainees were identified and tested for construct validity. Several methods were combined for assessing the properties using one framework. The results highlighted that inter-rater reliability was evident in twenty of the tools. As numerous tools are available the value of auditing inter-rater reliability between tools to demonstrate the individual construct validity is evident.

Hills (et al. 2009) assessed the National Institutes of Health Stroke Scale, a measure of neurological function, which yields multiple correct answers. Over 7,000 certified raters completed two standard videotaped patient examinations. Allowing multiple answers resulted in a greater variability of test results by examiners. The researchers suggested deriving a set of best answers through expert consensus where possible and then teaching raters how to derive these answers using a required interactive training module.

A study of observer atypia grading in endoscopic ultrasound demonstrated that even using governing guidelines a highly substantial variability is possible (Arville et al. 2014). Three participants, all board certified, received training. Randolph free marginal multirater kappa (RFMMK) was used to evaluate inter-observer variability. A RFMMK result of 0.18 was recorded. The assessment highlights that variability is possible despite following the Papanicolaou Society of Cytopathology's grading criteria.

DeFries (et al. 2011) investigated rater drift and maintaining rater integrity by examining changes over time in assessment of psychiatric disorders. 167 raters were trained on a standardised training protocol. Post training, videotaped interviews of patients were scored. To assess rater drift, data from this initial test was compared to data captured twelve months later at a refresher session. The study found that rater drift had occurred over the twelve month period and resulted in a lower overall intraclass coefficient. The researchers suggest that without ongoing training, a lack of consistency or rater drift over time will decrease inter-rater reliability and assessment integrity. Periodic training may help to prevent rater drift from in the assessment of health care accreditation standards.

Nisbet (et al. 2011) examined a reduction in inter-rater variability in the assessment of nuchal translucency image quality. Inter-rater agreement was evaluated before and after an intervention where the assessors were required to refer to a detailed resources manual designed to reduce the subjectivity inherent in image assessment. After the intervention, inter-rater agreement levels generally remained moderate. The study highlighted areas of image assessment that required critical review. The researchers recommended audit bodies regularly review inter-rater agreement to ensure consistency.

Another study audited the inter-observer agreement between primary graders and an expert grader in a diabetic retinopathy-screening program (Patra et al. 2009). For inter-grader assessments, an expert grader blinded to the primary results independently reviewed the images. The expert grader was an experienced ophthalmologist with a special interest in diabetic eye disease. The primary graders and expert inter-observer's agreement overall was 85% demonstrating an acceptable level of accuracy of grading in the program. The audit recommends performing a re-audit on the inter-rater reliability every twelve months for quality assurance.

Craddock (et al. 2010) evaluated the need to assess the reliability of quality development tools used to assess structured self management education programs against agreed standards. Eight newly designed programs were reviewed by eight independent assessors against agreed standards using the quality development assessment tools. The results showed 27% of assessments had poor reliability in determining whether the program met the agreed standards or not. The study highlights that even well documented quality development tools and processes have the potential to be unreliable. When there are multiple assessors, inter-rater reliability needs to be tested to assure standards are being met.

6.0 Findings & Limitations

The outcome of the literature review and analysis process has found similar results to the previous review in 2007 on the intra-rater and inter-rater reliability in health care accreditation (Greenfield et al. 2007). That is, there is insufficient empirical research into best practices to ensure reliability of accreditation surveyors of health care organisations. The majority of the literature available uses statistical analysis to evaluate inter-assessor reliability and is focused predominantly on clinical care. The results of the literature review have also shown that since 2007 there have only been a few key authors contributing to research on the evaluation of inter-assessor reliability in health service accreditation, which may have the effect of a higher susceptibility of bias. The limited depth of contributors exemplifies the deficiency of contemporary research.

However, for accreditation, assessors need to be able to make reliable judgments from survey to survey for grading to have legitimacy. (Shaw et al. 2010; Greenfield et al. 2009). The aim of this review was specifically to identify evidence for strategies that may improve the inter-assessor reliability during accreditation to ensure that there is a consistent assessment of organisations to the NSQHS Standards and to promote confidence in the Australian Health Service Safety and Quality Accreditation Scheme.

As there is limited literature specifically relating to accrediting health care organisation studies were included across a diverse range of disciplines or contexts. It has been argued that this will help to achieve a higher level of abstraction and that results can be generalisable (Thomas & Harden 2008). However, there is also the possibility that combining research from multiple disciplines may potentially de-contextualise the findings and the assumption that the concepts identified in one setting are applicable to the other may not be valid (Sandelowski & Barroso, cited in Thomas & Harden 2008, p. 2). By preserving the context of identified studies in this thematic analysis and providing summaries, detailing the methods, setting and sample, readers can self-evaluate their applicability to health care accreditation.

Five broad themes were identified within the literature as impacting on inter-assessor reliability: workforce management and assessor selection; assessor experience and personality; training and continued professional development; assessor teams; and regular auditing.

A number of studies suggested that inter-assessor reliability could be improved with workforce management, with a focus on assessor selection, providing appropriate leadership, balancing workforce demands, allowing adequate preparation time and statistically evaluating individual rater competency and reliability (Greenfield et al. 2009; Greenfield et al. 2012; Hinchcliff et al. 2013; Kott et al. 2008). Considering whether to publish health care organisation accreditation results may also impact on inter-assessor reliability and workforce management. Accreditation results remain confidential in Canada and as a result assessors operate in the capacity of a counsellor, whereas in France surveyors are commonly seen as inspectors with the result requiring greater public transparency and accountability (Touati & Pomey 2009). In Australia, clarifying the role and perceived style for the assessor workforce, which may also be influenced by public disclosure, will likely impact inter-assessor reliability.

Four papers, evaluating only a small sample of participants, showed that experienced raters were more likely to correctly and reliably score assessments (Kulkarni, Walker & Carter 2014; Olsen et al. 2014; Kerl et al. 2012). A larger study (Miller et al. 2011) reported the influence of personality traits on inter-assessor reliability and scoring tendencies. In instances where assessor experience level was shown to be of limited significance, some form of training had previously been provided (Beckman et al. 2014; Kott & Swartz 2012; Chari et al. 2012; Smith et al. 2013).

Numerous studies postulate that education and training of assessors reduces inter-assessor variability. Assessor training was shown to promote interpretation consistency and reliability across multiple disciplines (Doughton et al. 2012; Garcia-Reyes et al. 2014; Greenfield et al. 2009; Greenfield et al. 2009; Iwarsson & Peterson 2012; Lee & Choi 2014; Quigg & Lado 2009). Gaur's (et al. 2010) study makes it apparent that the modality and delivery of training may impact on inter-assessor reliability. Despite the provision of training, Brixely, Guse & Gorelick (2010) demonstrated that trained community assessors, without relevant professional experience, are less likely to produce a high level of inter-assessor reliability. A greater emphasis on training for particular standards or assessment criteria may improve areas of consistently low inter-assessor reliability (Abdel Baki et al. 2012; Yao, Foster & Aldrich 2009). However, a continuous

poor inter-assessor reliability may be an indicator of a poor process or a system not correctable through education (Bell, Aldinger & Richey 2011).

The use of assessor teams of two or three to achieve a high degree of agreement was a common theme in the literature. Teams achieve a higher level of reliability often facilitated through discussion to arrive at consensus (Yao, Foster & Aldrich 2009; Schildmeijer et al. 2014; Greenfield et al. 2013). However, composition and type of staff in review teams must be considered as multi-discipline teams typically recorded a lower inter-assessor reliability than teams of peers with mixed experience levels. (Hutchinson et al. 2010).

Despite individual accreditation programs in different countries having similarities in methodology, the content of standards and assessment procedures are quite diverse (Shaw et al. 2010). As such there is the need to audit assessment tools and accreditation processes to address identified issues. Several studies emphasised the need to review inter-assessor reliability to ensure that standards are being met (Greenfield et al. 2011; Abdell Baki et al. 2012; Carriker & Isaacs 2013; Carson, Fitch & Vachon 2009; Jelovsek, Kow & Diwadkar 2013; Hills et al. 2009; Arville et al. 2014; Nisbet et al. 2011; Craddock et al. 2010). Furthermore, repeated auditing showed a continued improvement in reliability (DeFries et al. 2011; Patra et al. 2009). There are a number of methods to statistically assess reliability, including correlation tests like Pearson, Kendall's tau, intraclass correlation coefficient and Spearman's rho, and reliability tests like Cronbach's alpha (Tan et al. 2014). These reliability and correlation tests are able to evaluate the consistency of scores from different assessors to the same subject (Liao, Hunt & Chen 2010) and may in turn prompt targeted education and training.

6.1 Further Research

There remains limited research, specific to health care accreditation, for strategies that could improve inter-assessor reliability. Future research should focus on optimising the delivery of strategic training and continuing professional development interventions to improve inter-assessor reliability, as well as research to determine appropriate education content. Other areas of research interest may involve evaluating assessor-drift over time of assessor interpretations and the examination of personal bias on surveyor assessments.

7.0 Conclusion

Promoting inter-assessor reliability is crucial to ensure health professional and managers remain engaged in accreditation programs (Hinchcliff et al. 2013). Five key themes identified in this analysis of the literature from a diverse range of disciplines signify focus areas to target as improvement opportunities for the maximisation of inter-assessor reliability in health care accreditation. Analysis of the results of studies in each thematic area has further identified evidence that inter-assessor reliability could be promoted in health care accreditation through sufficient assessor workforce management, selecting proficient experienced assessors, providing appropriate training and education in assessment criteria, promoting team surveying and conducting regular audits. The implementation of these best practice strategies in health care accreditation could drive the consistent application of NSQHS Standards & promote confidence in the Australian

Health Service Safety and Quality Accreditation (AHSSQA) Scheme. Further research, specifically in health care accreditation, is required to test and develop the different strategies.

8.0 Appendices

8.1 Appendix 1

National Safety and Quality Health Service Standards:

1. Governance for Safety and Quality in Health Service Organisations
2. Partnering with Consumers
3. Preventing and Controlling Health care Associated Infections
4. Medication Safety
5. Patient Identification and Procedure Matching
6. Clinical Handover
7. Blood and Blood Products
8. Preventing and Managing Pressure Injuries
9. Recognising and Responding to Clinical Deterioration in Acute Health Care
10. Preventing Falls and Harm from Falls

(ACSQHC 2011)

8.2 Appendix 2

The Australia Health Service Safety and Quality Accreditation Scheme approved accrediting agencies:

1. Australian Council on Healthcare Standards
2. National Association of Testing Authorities (NATA), Australia
3. BSI Group ANZ Pty Ltd
4. Quality Innovation Performance Pty Ltd
5. Global Mark P/L
6. HDAA Pty Ltd
7. SAI Global
8. Total Quality Certification Services International (TQCSI)
9. Institute for Healthy Communities (IHCA)
10. International Standards Certifications

(ACSQHC 2014)

9.0 References

Abdel Baki, S. G., Weedon, J., Arnedo, V., Chari, G., Koziorynska, E., Lushbough, C. A., . . . Grant, A. C. (2012). An exploration of the diagnostic reliability of EEG using latent class analysis. *Epilepsy Currents, 12*(1).

Adams, W. (2011). Changes to National Safety and Quality Health Service Standards: How Will It Impact Us in Our Day Surgery Facilities?. *Day Surgery Australia*, 10(1), 3.

Arville, B., Evans, H., Al Diffalha, S., Barkan, G., Pambuccian, S., & Atieh, M. (2014). Inter-observer variability of two-tiered atypia grading in endoscopic ultrasound guided fine needle aspirations of cystic mucinous pancreatic neoplasms. *Laboratory Investigation, 94*, 96A.

Australian Commission on Safety and Quality in Health Care (ACSQHC) 2011, *National Safety and Quality Health Service Standards*, ACSQHC, Sydney. Access <http://www.safetyandquality.gov.au/wp-content/uploads/2011/09/NSQHS-Standards-Sept-2012.pdf> [12 August 2014]

Australian Commission on Safety and Quality in Health Care (ACSQHC) 2012, *Guide to the application process for approval of Accrediting Agencies*. ACSQHC, Sydney. Access <http://www.safetyandquality.gov.au/wp-content/uploads/2012/03/Guide-to-the-application-process-for-the-approval-of-accrediting-agencies-September-2012.pdf> [04 August 2014]

Australian Commission on Safety and Quality in Health Care (ACSQHC) 2014, *Approved Accrediting Agencies*. ACSQHC, Sydney. Access <http://www.safetyandquality.gov.au/wp-content/uploads/2014/04/Contact-details-for-accrediting-agencies-approved-to-assess-health-service-organisations-against-the-NSQHS-Standards.pdf> [09 August 2014]

Baumeister, R. F., & Leary, M. R. (1997). Writing narrative literature reviews. *Review of general psychology*, *1*(3), 311.

Beckmann, A., Hamilton-Giachritsis, C., Johns, N., & Kenyon, S. (2014). The inter-rater reliability of the birmingham symptom specific obstetric triage system (BSOTS). *Archives of Disease in Childhood: Fetal and Neonatal Edition, 99*, A24-A25.

Bell, T., Aldinger, J., & Richey, H. (2011). A prospective evaluation of E/M coding variability. *Journal of Hospital Medicine, 6*(4), S12-S13.

Bohigas, L., Brooks, T., Donahue, T., Donaldson, B., Heidemann, E., Shaw, C., & Smith, D. (1998). A comparative analysis of surveyors from six hospital accreditation programmes and a consideration of the related management issues. *International Journal for Quality in Health Care*, *10*(1), 7-13.

Brixey, S. N., Guse, C. E., & Gorelick, M. (2010). Reliability and validity of child passenger safety restraint observations by community observers. *Traffic Inj Prev, 11*(6), 573-577. doi: 10.1080/15389588.2010.508083

Carriker, C., & Isaacs, P. (2013). Do you see what i see? A validation process for an observational hand hygiene compliance program that works. *Am J Infect Control, 41*(6), S55-S56.

Carson, M. G., Fitch, M. I., & Vachon, M. L. S. (2009). Measuring patient outcomes in palliative care: a reliability and validity study of the Support Team Assessment Schedule. *Palliative Medicine, 14*(1), 25-36.

Chari, G., Abdel-Baki, S. G., Omurtag, A., Weedon, J., Lushbough, C. A., Koziorynska, E., . . . Grant, A. C. (2012). Inter- and intra-rater reliability of EEG interpretation: A large, single-center study. *Epilepsia, 53*, 130.

Cradock, S., Stribling, B., Dallosso, H. B., Daly, H., Carey, M. E., Cullen, M., . . . Yates, T. (2010). The need for assessing reliability of Quality Development (QD) tools in Structured Self Management Education (SSME) programmes in diabetes. *Diabetic Medicine, 27*(2), 119.

DeFries, A., Rothman, B., Yavorsky, C., Opler, M., Gordon, J., Tourian, K., . . . Vialet, C. (2011). Quantifying rater drift on the HAM-D: Implications for reliability, sample size, and ongoing training strategy. *European Neuropsychopharmacology, 21*, S357-S358.

Doughton, J., Foley, H., Morrison, S., Plank, A., Fay, M., & Martin, J. (2012). Reducing variability in rectal cancer target volume delineation - An education intervention. *J Med Imaging Radiat Oncol, 56*, 258.

Garcia-Reyes, K., Palmeri, M. L., Kauffman, C. R., Polascik, T. J., & Gupta, R. T. (2014). Detection of prostate cancer with multiparametric MRI (mpMRI): Effect of dedicated reader education on accuracy/confidence of index and anterior cancer diagnosis. *Abdom Imaging, 39*(3), 668.

Gaur, R., Kaviani, H., Bansal, S., & Lee, N. (2010). Rater training on PANSS and SANS. *Schizophr Res, 117*(2-3), 265-266.
Greenfield, D., & Braithwaite, J. (2008). Health sector accreditation research: a systematic review. *International Journal for Quality in Health Care*, 20(3), 172-183.
Greenfield, D., & Braithwaite, J. (2009). Developing the evidence base for accreditation of healthcare organisations: a call for transparency and innovation. *Qual Saf Health Care, 18*(3), 162-163. doi: 10.1136/qshc.2009.032359
Greenfield, D., Braithwaite, J., & Pawsey, M. (2008). Health care accreditation surveyor styles typology. *Int J Health Care Qual Assur, 21*(5), 435-443. doi: 10.1108/09526860810890422
Greenfield, D., Hinchcliff, R., Westbrook, M., Jones, D., Low, L., Johnston, B., ... & Braithwaite, J. (2012). An empirical test of accreditation patient journey surveys: randomized trial. *International Journal for Quality in Health Care*, 24(5), 495-500.
Greenfield, D., Moldovan, M., Westbrook, M., Jones, D., Low, L., Johnston, B., ... & Braithwaite, J. (2011). An empirical test of short notice surveys in two accreditation programmes. *International Journal for Quality in Health Care*, 24 (1), 65-71.
Greenfield, D., Pawsey, M., & Braithwaite, J. (2011). What motivates professionals to engage in the accreditation of healthcare organizations?. *International Journal for Quality in Health Care*, *23*(1), 8-14.
Greenfield, D., Pawsey, M., Naylor, J., & Braithwaite, J. (2009). Are accreditation surveys reliable? *Int J Health Care Qual Assur, 22*(2), 105-116. doi: 10.1108/09526860910944601
Greenfield, D., Pawsey, M., Naylor, J., & Braithwaite, J. (2013). Researching the reliability of accreditation survey teams: lessons learnt when things went awry. *Health Information Management Journal, 42*(1), 4-10.
Greenfield, D., Travaglia, J., Pawsey, M., & Braithwaite, J. (2007). Intra-Rater And Inter- Rater Reliability In Health Care Accreditation: Literature Review. *Centre For Clinical Governance Research (Unsw)*.
Hills, N. K., Josephson, S. A., Lyden, P. D., & Johnston, S. C. (2009). Is the NIHSS certification process too lenient? *Cerebrovasc Dis, 27*(5), 426-432. doi: 10.1159/000209237
Hinchcliff, R., Greenfield, D., Moldovan, M., Westbrook, J. I., Pawsey, M., Mumford, V., & Braithwaite, J. (2012). Narrative synthesis of health service accreditation literature. *BMJ quality & safety*, bmjqs-2012.
Hinchcliff, R., Greenfield, D., Westbrook, J. I., Pawsey, M., Mumford, V., & Braithwaite, J. (2013). Stakeholder perspectives on implementing accreditation programs: a qualitative study of enabling factors. *BMC Health Serv Res, 13*, 437. doi: 10.1186/1472-6963-13-437
Hurst, K. (1997). The nature and value of small and community hospital accreditation. *International Journal of Health Care Quality Assurance*, *10*(3), 94-106.
Hutchinson, A., Coster, J. E., Cooper, K. L., McIntosh, A., Walters, S. J., Bath, P. A., . . . Ratcliffe, J. (2010). Comparison of case note review methods for evaluating quality and safety in health care. *Health Technology Assessment, 14*(10), 1-170.
International Society for Quality in Health Care (ISQua). 2012. Who we are? Access <http://www.isqua.org/who-we-are/who-we-are> [06 August 2014]
Iwarsson, J., & Reinholt Petersen, N. (2012). Effects of consensus training on the reliability of auditory perceptual ratings of voice quality. *J Voice, 26*(3), 304-312. doi: 10.1016/j.jvoice.2011.06.003
Jayasinghe, U. W., Marsh, H. W., & Bond, N. (2006). A new reader trial approach to peer review in funding research grants: An Australian experiment. *Scientometrics*, *69*(3), 591-606.
Jelovsek, J. E., Kow, N., & Diwadkar, G. B. (2013). Tools for the direct observation and assessment of psychomotor skills in medical trainees: a systematic review. *Med Educ, 47*(7), 650-673. doi: 10.1111/medu.12220
Joint Accreditation System of Australia and New Zealand (JAS-ANZ), 2012. About Us. Access <http://www.jas-anz.org/index.php?option=com_content&task=blogcategory&id=13&Itemid=32> [05 August 2014]
Kerl, J. M., Schoepf, U. J., Bauer, R. W., Tekin, T., Costello, P., Vogl, T. J., & Herzog, C. (2012). 64-slice multidetector-row computed tomography in the diagnosis of coronary artery disease: interobserver agreement among radiologists with varied levels of experience on a per-patient and per-segment basis. *J Thorac Imaging, 27*(1), 29-35. doi: 10.1097/RTI.0b013e3181f82805
Kott, A., Spear, C., Miller, D., Butler, A., & Markovic, O. (2008). Proven methodology for selecting

raters. *European Neuropsychopharmacology, 18*(S4), S498.
Kott, A., & Swartz, J. (2012). Does rater experience affect UPDRS inter-rater reliability? *Movement Disorders, 27*, S100.
Kulkarni, U. M., Walker, W., & Carter, I. I. I. W. E. (2014). INTER-RATER RELIABILITY IN TBI CLASSIFICATION. *American Journal of Physical Medicine & Rehabilitation*, a94-95.
Lee, S., & Choi, J. I. (2014). Intra- and interobserver reliability of gray scale/dynamic range evaluation of ultrasonography using a standardized phantom. *33*(2), 91-97. doi: 10.14366/usg.13021
Liao, S. C., Hunt, E. A., & Chen, W. (2010). Comparison between inter-rater reliability and inter-rater agreement in performance assessment. *Ann Acad Med Singapore, 39*(8), 613-618.
Miller, A. K., Rufino, K. A., Boccaccini, M. T., Jackson, R. L., & Murrie, D. C. (2011). On individual differences in person perception: raters' personality traits relate to their psychopathy checklist-revised scoring tendencies. *Assessment, 18*(2), 253-260. doi: 10.1177/1073191111402460
Nisbet, D., McLennan, A., Robertson, A., Schluter, P. J., & Hyett, J. (2011). Reducing inter-rater variability in the assessment of nuchal translucency image quality. *Fetal Diagn Ther, 30*(2), 128-134. doi: 10.1159/000326339
Olsen, E., Dunkel, B., Barker, W. H., Finding, E. J., Perkins, J. D., Witte, T. H., . . . Piercy, R. J. (2014). Rater agreement on gait assessment during neurologic examination of horses. *J Vet Intern Med, 28*(2), 630-638. doi: 10.1111/jvim.12320
Patra, S., Gomm, E. M. W., Macipe, M., & Bailey, C. (2009). Interobserver agreement between primary graders and an expert grader in the Bristol and Weston diabetic retinopathy screening programme: a quality assurance audit. *Diabetic Medicine, 26*(8), 820-823. doi: http://dx.doi.org/10.1111/j.1464-5491.2009.02767.x
Quigg, M., & Lado, F. A. (2009). Interrater reliability to assure valid content in peer review of CME-accredited presentations. *J Contin Educ Health Prof, 29*(4), 242-245. doi: 10.1002/chp.20042
Sandelowski M, Barroso J: Handbook for Synthesising Qualitative Research New York: Springer; 2007.
Schildmeijer, K., Nilsson, L., Årestedt, K., & Perk, J. (2012). Assessment of adverse events in medical care: lack of consistency between experienced teams using the global trigger tool. *BMJ Quality & Safety, 21*(4), 307-314. doi: 10.1136/bmjqs-2011-000279
Shaw, C., Groene, O., Mora, N., & Sunol, R. (2010). Accreditation and ISO certification: do they explain differences in quality management in European hospitals?. *International Journal for Quality in Health Care*, 22(6), 445-451
Siggins Miller. (2009). Surveyor participation in safety and quality accreditation. A report to the Australian Commission on Safety and Quality in Health Care.
Smith, C. A., Chimera, N. J., Wright, N. J., & Warren, M. (2013). Interrater and intrarater reliability of the functional movement screen. *J Strength Cond Res, 27*(4), 982-987. doi: 10.1519/JSC.0b013e3182606df2
Tan, J., Liu, H., Leyden, J. J., & Leoni, M. J. (2014). Reliability of Clinician Erythema Assessment grading scale. *J Am Acad Dermatol, 71*(4), 760-763. doi: 10.1016/j.jaad.2014.05.044
Thomas, J., & Harden, A. (2008). Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC medical research methodology*, *8*(1), 45.
Touati, N., & Pomey, M. P. (2009). Accreditation at a crossroads: are we on the right track? *Health Policy, 90*(2-3), 156-165. doi: 10.1016/j.healthpol.2008.09.007
Wimer, J. W. (2007). An Investigation of Interrater Reliability Among Athletic Training Accreditation Site Visitors. *Journal of Allied Health*.
Yao, Y., Foster, K., & Aldrich, J. (2009). Interrater Reliability of a Team-Scored Electronic Portfolio. *Journal of Technology and Teacher Education, 17*(2), 253.