

**Development and validation of a visual grading scale for assessing image quality
of AP pelvis radiographic images**

Hussien Abid Ali Bakir Mraity^{1,2} MSc, PhD.

Andrew England² MSc, PhD.

Simon Cassidy² MSc, PhD.

Peter Eachus² MSc, PhD.

Alejandro Dominguez^{3,4} BSc, MSc.

Peter Hogg^{2,5}, BSc, MPhil

1. University of Kufa, Faculty of Science, Department of Physics, Najaf, Iraq.
2. School of Health Sciences, University of Salford, United Kingdom.
3. Haute Ecole de Sante Vaud, University of Applied Sciences and Arts Western Switzerland, Lausanne.
4. School of Health Sciences, University Center of Legal Medicine, Switzerland.
5. Department of Radiography, Karolinska Institute, Stockholm, Sweden.

ABSTRACT

Objective: Apply psychometric theory to develop and validate a visual grading scale for assessing visual perception of AP pelvis digital image quality.

METHODS: Psychometric theory was used to guide scale development. Seven phantom and 7 cadaver images of visually and objectively predetermined quality were used to help assess scale reliability and validity. 151 volunteers scored phantom images; 184 volunteers scored cadaver images. Factor analysis and Cronbach's alpha were used to assess scale validity and reliability.

RESULTS: A 24 item scale was produced. Aggregated mean volunteer scores for each image correlated with the rank order of the visually and objectively predetermined image qualities. Scale items had good inter-item correlation (≥ 0.2) and high factor loadings (≥ 0.3). Cronbach's alpha (reliability) revealed that the scale has acceptable levels of internal reliability for both phantom and cadaver images ($\alpha = 0.8$ and 0.9 , respectively). Factor analysis suggested the scale is multidimensional (assessing multiple quality themes).

CONCLUSION: This study represents the first full development and validation of a visual image quality scale using psychometric theory. It is likely that this scale will have clinical, training and research applications.

ADVANCES IN KNOWLEDGE: This article presents data to create and validate visual grading scales for radiographic examinations. The visual grading scale, for AP pelvis examinations, can act as a validated tool for future research, teaching and clinical evaluations of image quality.

Introduction

The reliability and validity of image quality assessment methods are crucial in medical imaging. Reliability refers to precision of a given assessment, and therefore high reliability demands a small amount of random error. Validity represents the ability of measures to define the phenomenon, and a high validity necessitates low systematic error (1). In medical imaging, image quality is a phenomenon of considerable complexity resulting from the diversity of radiographic projections and the ongoing improvement in radiologic technology (2). As such it is easy to see the difficulties in identifying an image quality measure that has high validity and reliability. A wide range of quality assessment methods are currently used in medical imaging practice (1,3).

Several methods have been described to assess image quality (4), these include physical (e.g. signal to noise ratio (SNR), contrast to noise ratio (CNR), modulation transfer function (MTF), detective quantum efficiency (DQE)), visual grading and observer performance (e.g. receiver operating characteristic (ROC)). SNR is a surrogate measure of feature contrast and MTF is a quantitative measure of a system's resolution (5). These highly precise physical metrics relate to a measure of detectability of relevant features but they do not measure the visibility/reproduction of features directly (1). In this context, MTF and contrast can be measured very precisely, but observers may not directly observe MTF or contrast as entities since they almost always observe certain features (e.g. anatomy) within an image. The observer must then decide on the clinical relevance of these features, which is the difficult part of the assessment task (6). By contrast, well designed and controlled observer performance studies provide reliable and clinically useful information. Visual grading methods can provide important information about clinical image quality, however for a number of reasons they are the least reliable of all image evaluation methods (7). To illustrate, with this approach the anatomical structures under evaluation must be pre-specified; no formal and validated data on this exists, and it is likely that these will be highly variable between publications and this makes comparisons difficult (8).

Poor reliability of visual grading methods may arise from a lack of universally accepted standards in terminology/criteria, and the lack of validated scales within the literature

(9,10). Image quality assessments are, therefore, almost always subjective; this in turn leads to inter and intra observer variability which is a common problem in medical imaging (11). The only published criteria for visual image quality assessment was provided by the Commission of European communities (CEC) in 1996 (12). However, no robust evidence was published alongside the CEC criteria to validate them. They were produced in an era of film, so many of the criteria need revising to fit with the digital environment (10, 13, 14), while other important criteria are missing (e.g. iliac crests in AP pelvis radiography) (15). Revising the existing criteria is essential, new digital concepts have been established such as signal, noise and signal to noise ratio and old film-based concepts such as darkening and blackness have been eliminated (16,17). In an attempt to address the current problems in visual image quality assessment, Mraity et al (10,15) has previously piloted a novel approach to create and validate an image quality visual grading scale. This approach is based on psychometric theory (also known by psychometrics). Psychometrics is a branch of psychology concerned with the objective measurement of attributes that cannot be measured directly such as image perception (18). It is suggested here that the application of psychometric theory is likely to advance efforts at obtaining valid and reliable measurements of an observer's perception, since the evaluation of image quality involves the interaction between observer perception and image information. The successful pilot by Mraity and colleagues developed and validated a visual grading scale for assessing the image quality of PA chest images (10).

In order to build on the previous work by Mraity, our paper uses AP pelvis radiography as a model in order to provide a full and detailed development and validation study of a visual grading scale for measuring observer perception of image quality (15). The rationale for selecting the AP pelvis projection was that the pelvic region includes the reproductive organs which means they are inevitably exposed to primary beam (19, 20); also pelvis / hip radiography has been reported to be the third most frequent examination in the UK (21).

Further rationale for our work is based on literary analysis, indicating that there are no standard approaches for assessing [AP pelvis] image quality (15). Some publications

(e.g. 22) report on the measurement of noise and contrast, others use CEC criteria (e.g. 23). In some cases the CEC criteria are actively avoided and 'diagnostic usefulness', reflecting noise and image clarity, have been used instead (24). Two recent studies (25, 26) attempted to develop an optimisation framework for AP pelvis. One of these (25) used CEC criteria with the addition of three more criteria; the other (26) used the CEC criteria verbatim. Despite an intensive literature search no validated visual grading scale for image quality assessment of AP pelvis images was identified. It is clear that a complete lack of standards exist for assessing visual perception of AP pelvis image quality (15).

The aim of this paper was to apply psychometric theory, in particular classical test theory, in a novel fashion to construct and validate a visual grading scale for assessing the visual perception of image quality of AP pelvis.

Method

Ethical approval was obtained from the University of Salford in order to recruit volunteers for image quality assessments, membership of the focus group and also to use images acquired from human cadavers. Our method followed the pilot work described by Mraity et al (2014), which is based upon the principles of psychometric scale development and validation (15, 27, and 28). The validation phases took place in a room with PCs and computer screens dedicated to medical image analysis. Initial scale validation was conducted using phantom images, as phantoms were readily available. Further validation was then conducted using cadaver images, so that a range of pelvis sizes and shapes could be assessed. Four cadavers without pathology were used to acquire images of different estimated image qualities.

Scale development and validation

Two publications were identified which outlined how psychometric theory could be applied to develop and validate a scale for assessing perception of image quality (10, 15); similar notable applications of this theoretical perspective include the development of perceived self-efficacy scales by Bandura (28), who's guidelines have been widely

used to develop self-efficacy scales in many different disciplines (e.g. health and education sectors).

Development and validation comprised of five steps:-

1. Define AP pelvis image quality; this included identifying the main factors/themes of image quality on pelvic images. Establishing a clear definition of the construct leads to the generation of a pool of items that represents the construct (27, 29). The construct normally represents different factors/themes of image quality, when combined they make up the whole construct. Two examples of themes for AP pelvis could be 'anatomical detail' and 'technical detail'.
2. Create the scale; this included generating the scale items from each theme identified in point 1. Scale items should cover the entire construct (i.e. image quality of AP pelvis), which is important for scale validity and reliability. By definition, an item is a statement that refers to a specific image quality attribute. Identification of potential items was conducted initially through literature analysis. This analysis was conducted by the researchers using relevant journal literature, textbooks and reports (e.g. CEC guidelines (12)). Based on identified themes, the image quality attributes within each theme were classified and written within a draft document.
3. Assess scale face validity using a panel of volunteers with relevant insight. In this step items were created and then reviewed critically by a focus group (6 radiographers, 1 radiologist and 1 medical physicist) to assess face validity. Face validity can be defined as the degree to which the scale is subjectively viewed as covering the concept (i.e. image quality) it is intended to measure. Focus group analysis clarified item wording and the number of items. From this, an initial scale of 29 items was created. Some scale items were worded negatively (i.e. items 9, 15, 24, see Appendix) and others positively to minimise affirmation bias.
4. Volunteer observers use the newly drafted scale to score images of 'estimated image qualities'. A 5 point Likert scale was used for scoring (1 = strongly disagree; 5 =strongly agree).

5. Analyse volunteer observer data, using appropriate statistical tests (27, 29). It should be noted that the final item selection will be based upon standard psychometric principles and statistical testing (27, 30, 31). These statistical tests will be described later in this methods section.

Estimation of image quality

Seven phantom and 7 cadaver images, of estimated and different qualities ranging from low to high, were produced to validate the scale. Image production was achieved by manipulating kVp and mAs; kVp affects both photon density and beam quality whereas mAs affects photon density only, which in turn impacts on the image quality (32). It is worth noting that the high quality image was based on standard acquisition parameters including tube current termination using automatic exposure control (AEC). Image selection and quality ranking was based on signal to noise ratio (SNR) (1, 33) by way of an objective image quality measure *and* consensus opinion (6 experienced radiographers, 1 radiologist and 1 medical physicist).

Seven phantom images were used for the initial scale validation. An opaque anthropomorphic pelvis phantom (Alderson, 19 cm thick) was positioned using the standard AP radiographic positioning (37). For the phantom images, a series of experiments were conducted to generate images, ranked from low to high SNR. The SNR was calculated using four regions of interest in order to obtain an overall objective measure (Figure 1). SNR calculations were conducted using the following equation:

$$\frac{\text{Mean signal (taken from ROI}_{1, 2, 3, \text{ and } 4})}{\text{Noise}(\sigma \text{ of ROI}_5)} \dots\dots(1)$$

The mean signal value was obtained by averaging the signals of the four ROIs, whereas the noise value was identified from the standard deviation (σ) of the background ROI. Then by dividing the mean signal value by noise value the SNR was calculated.

The 7 phantom images were then presented to the focus group for visual inspection to determine whether they agreed that the images were different and ranked appropriately (high to low as per SNR ranking, see Figure 2). The focus group used definitions (Table

1) from various sources to help identify image features which were consistent with the most common disorders that pelvis radiology is tasked with investigating (34, 35, 36).

Table 1. Examples of the definitions used to help identify the image with different and predefined quality (34,35)

Image quality classification	Definition
Uninterpretable image	This kind of image should be repeated
Non-diagnostic quality	Lack or no clinically useful information for answering the primary clinical question
Limited quality	Acceptable, with some technical defect but still adequate for diagnostic purposes
Diagnostic quality	Acceptable, with no technical defect likely to impair using the image for diagnosis
Exemplary quality	Good, more than adequate for diagnostic purposes

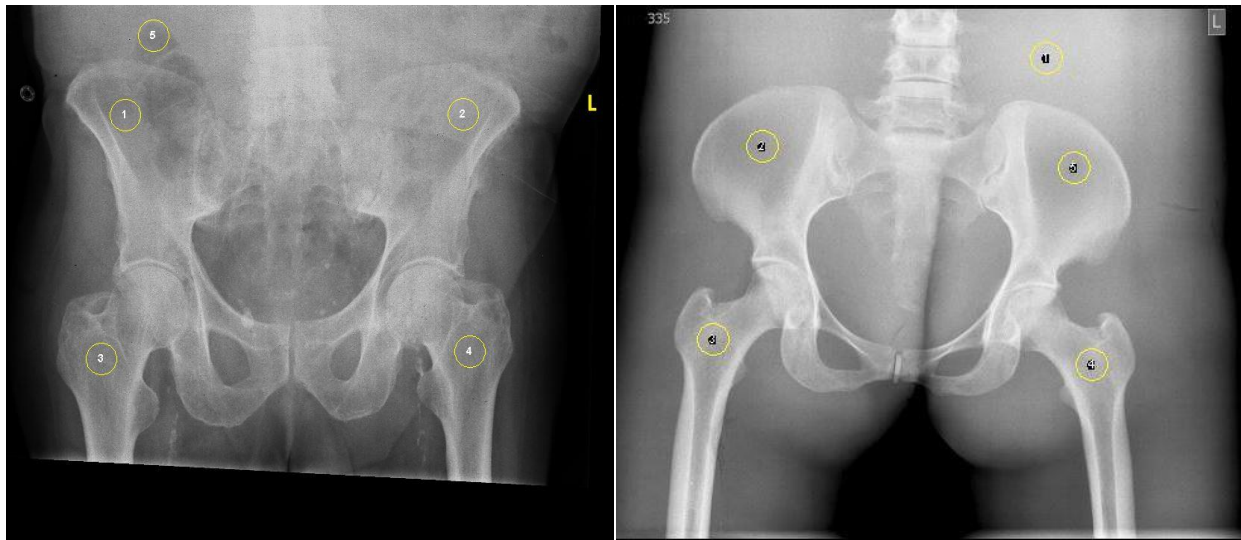


Figure 1. This figure illustrates the four ROIs used to calculate the mean signal values across the pelvis and the fifth ROI (uniform area) used to identify noise (σ) for the purpose of SNR calculation in both phantom and cadaver images (Left, cadaver image; right, phantom image)

The cadaver images were used for further validation of the scale. Unlike phantom images, cadaver images can demonstrate anatomical variance, structured noise, and

soft tissue shadow variations. Acquisition factors which led to the acquisition of the final seven phantom images were used as guidance to acquire the 7 cadaver images. Again, it is worth noting that the high quality image was based on standard acquisition parameters including tube current termination using an AEC. Similar to the phantom, the cadavers were positioned using standard radiographic positioning (37). Seven images of different SNRs were obtained. Additionally, for the 7 cadaver images, positional and technical errors were introduced by either rotating the cadaver body to either side, and/or through incorrect collimation. The reason behind these errors was to test the capability of the scale in characterising the effect of the positional errors in addition to the noise's effect on image quality. Similar to the phantom images, the cadaveric images were introduced to the focus group to ensure they were different in terms of quality and again ranked appropriately (high to low, see Figure 3). Again, this process was conducted with the aid of subjective definitions indicated in Table 1 (34, 35).

Phantom images were acquired on a Wolverton Acroma X-ray unit (high frequency generator with VARIAN 130 HS standard X-ray tube with a total filtration of 3 mm Aluminium equivalent). The same 35 cm x 43 cm Agfa image receptor and image processing was undertaken using an Agfa 35-X digitiser. All equipment quality control met the required specifications of Institute of Physics and Engineering in Medicine (IPEM) report 91 (38).

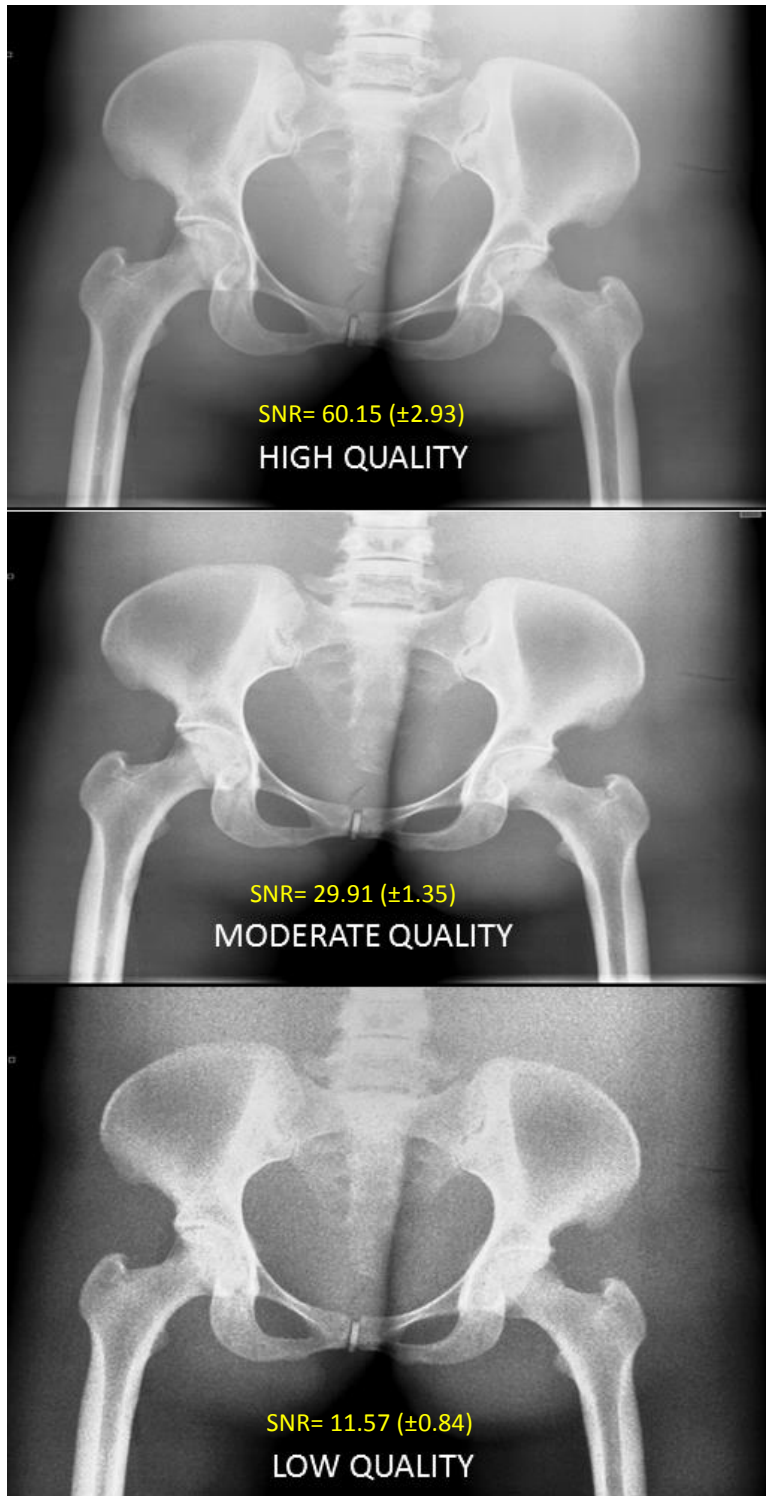


Figure 2. Examples of phantom images



Figure 3. Examples of cadaver images

Volunteers used for scale validation

Scale validation was conducted in two phases, first using phantom images and then using cadaver images. Scoring involved volunteers completing one scale for each image under controlled conditions: room lighting was dimmed (20-38 Lux) (39); 2.3 mega-pixel screens (22 inch liyama ProLite liquid crystal display monitors -B2206WS), calibrated to the DICOM greyscale standard. No image manipulation was permitted by the volunteers and they were blinded to image acquisition factors. Images were presented in a random order. Immediately prior to image scoring all volunteers received a training exercise in image quality. The training exercise focused on defining image quality, physical aspects of quality (e.g. contrast) and how acquisition factors can affect the quality of an image.

Sample size is an important aspect of psychometric scale validation. Reliability assessment and factor analysis requires a minimum sample size of at least 150 participants (27, 40, 41). Consequently, 151 volunteers (resulting in $7 \times 151 = 1057$ completed scales) scored the phantom images and 184 (resulting in $7 \times 184 = 1228$ completed scales) scored the cadaver images. Volunteers comprised qualified radiographers and student radiographers from six higher education institutions (Table 2). It was accepted that student radiographers may have lower levels of experience, however to achieve the required sample size they represented an easily accessed group. Also, it is important to note they had all received training about image quality appraisal in their undergraduate studies, and additional training for this research.

Image set	Radiography students			Qualified radiographers	Overall sample size
	Year 1	Year 2	Year 3		
Phantom	42	43	41	25	151
Cadaver	55	59	41	29	184

There was a small of overlap between the phantom and cadaver volunteer groups, 20% of those who participated in the phantom study also participated in the cadaver study.

Statistical testing

Different statistical tests were used to investigate outliers, testing the quality of individual items, items' correlations and their reliability, items pattern correlation and sample adequacy ready for factor analysis. Details of the statistics are indicated below.

Item analysis

Item analysis was conducted for each image ready for internal reliability testing and factor analysis. Scale items with a standard deviation (SD) >1.5 were excluded (42). Skewness for each item was also investigated; items with highly skewed scores (i.e. >+1 or < -1) were removed (31, 43). Correlation of scale items was examined; this was based on inter-item correlation coefficients among all items within a correlation matrix. Scale items were eliminated if more than half of their responses had an inter-item correlation value ≤ 0.2 (44).

Internal consistency

Internal consistency of scale items was assessed using Cronbach's Alpha coefficient (45). A value of 0.6 was used as a standard lenient cut off point for each extracted factor (46). Cronbach's Alpha was calculated for all images before and after removing the redundant items to see how this affected scale internal reliability. Redundant items are those items which had high skewed scores or had high standard deviations.

Factor analysis

Factor structure refers to the number and nature of factors as assessed by the items. A factor on its own represents a group of items which represent different image quality attributes. A factor (group of items) is assumed to have a specific image quality theme. Factor analysis was used to explore whether any group of items had an underlying factor structure, and to explore whether the scale is uni- or multi-dimensional. Scale dimensionality expresses whether the scale is assessing a single image quality theme (e.g. anatomical features) or multiple themes (e.g. anatomical and technical) (27). Additionally, factor analysis can help remove items that have no meaningful contribution to the overall scale (low correlated items). This is based upon a correlation between

scale items scores and generated factors for each of the images. This kind of correlation is also known as factor loading (41). Factor analysis was conducted in two ways. First, principal component analysis (unrotated) to generally assess scale factorial structure. Second, a rotated factor analysis (varimax) approach helped determine how many image quality themes the scale had. Principal component analysis was conducted on the remaining items for each of the images. By contrast, factor analysis using the varimax function was conducted on the same group of items again to examine whether another factorial structure existed, and to reduce confusion/overlap that might arise from principal component analysis (one item loaded high into two factors) (47). This analysis was conducted separately for all images. Prior to conducting factor analysis, it was necessary to confirm that correlation patterns between items were relatively compact. For this purpose a Kaiser-Meyer-Olkin (KMO) statistic was used.

Results and analysis

Phantom data

The SNR values of the 7 phantom images and the aggregated image quality scores are demonstrated in Table 3 together with the corresponding acquisition factors.

Table 3. This table presents SNR values, initial image quality scale scores of the 7 phantom images along with the corresponding acquisition factors.									
Image rank	kVp	mAs	SID (cm)	SNR		Image quality scale score			
				Mean	SD	Mean	SD	95% CI for mean	
1	85	20	110	60.15	2.93	111.9	12.6	109.9	113.9
2	80	16	110	42.72	2.12	107.3	10.9	105.3	109.3
3	75	12	110	34.44	1.57	105.9	10.6	103.9	107.9
4	70	9	110	29.91	1.35	101.9	11.5	99.9	103.9
5	65	7.1	110	23.08	1.12	97.2	13.1	95.2	99.2
6	57	5	110	16.66	1.01	80.0	12.4	78.0	82.0
7	50	4	110	11.57	0.84	62.8	11.9	60.8	64.8

Rank 1 = very good image quality; Rank 7= very poor image quality; Rank was derived through SNR and focus group discussion. CI, confidence interval SD : standard deviation

Using Grubb's equation and a scatter plot (48), 6 outliers were detected across 4 image sets and in order to minimise error these were removed - image 2 (1 outlier), image 5 (2 outliers), image 6 (2 outliers) and image 7 (1 outlier).

Scale Reliability

Cronbach's Alpha values ranged from 0.803 to 0.913; this was after the redundant items had been removed. Removing these items had no substantial influence on scale reliability.

Factor analysis

To begin with, KMO values ranged from 0.823 to 0.888 (minimum acceptable ≥ 0.5) across the 7 images (49), this provides an indication that the correlations between items were compact. Principal component analysis revealed a series of single dominant factors with an eigenvalue ≥ 1 , for all seven images. The variance (%) that these factors accounted for ranged from 28.8 to 40.21%. This suggests that these factors explained the most important information regarding the quality of the AP pelvis images (29). Therefore, this would support the scale's construct validity.

For 5 out of 7 images the rotated factor analysis demonstrated the first 3 factors had similar and high variance (Table 4). For the remaining 2 image sets, the first factor in each accounted for the highest variance - 25% and 37%, respectively; whereas the second largest factors accounted for 12.9% and 11 % of total variance.

Table 4. An example of rotated factor analysis (image number 2)			
Factor number	Total variance	% of Variance	Cumulative %
1	2.847	16.749	16.749
2	2.781	16.360	33.108
3	2.635	15.500	48.608
4	1.512	8.892	57.500

Factors number 1, 2, and 3 from image sets 1, 2, 4, 5, and 6 can be considered as significant factors (50, 51, 52), because of their high variance (%) and the number of items that are highly loaded onto them (≥ 0.33). These factors were retained with their items. Two factors from image set 7 and 1 factor from image set 3 were retained based on the same criteria (50, 51).

Any item that had low factor loading (≤ 0.3) was removed because of its low correlation with the extracted factor(s) (53). Factors that did not have an adequate number of items were eliminated to improve psychometric properties of the scale. After analysis and item removal, 24 items were retained in the final scale. This 24 item scale (see Appendix) was used in the second phase of validation with the seven cadaver images.

Cadaver data

The details of the objective image quality metrics associated with the seven cadaver images together with aggregated scores from each volunteer and each image can be seen in Table (5).

Table 5. This table presents SNR values, initial image quality scale scores of the 7 cadaver images along with the corresponding acquisition factors.

Image rank	SNR		Noise (σ)	Image quality score			
	Mean	SD		Mean	SD	95% CI for mean	
1	49.48	4.56	3.303	97.5	9.5	96.1	98.9
2	30.74	2.55	3.400	89.9	10	88.5	91.3
3	25.62	6.12	5.824	82.3	9	81.0	83.6
4	24.53	2.88	4.693	81	10	79.6	82.4
5	20.71	2.50	5.866	74	9	72.7	75.3
6	14.91	1.03	7.089	76.3	10	74.9	77.7
7	12.51	1.65	9.915	63	11	61.4	64.6

Rank 1 = very good image quality; Rank 7= very poor image quality; Rank was derived through SNR and focus group discussion. CI, confidence interval SD : standard deviation

Cadaver scale data were analysed using the same approach used for phantom data. One outlier was detected from image set 1. This was removed from the data for the subsequent analysis to avoid systematic error.

Item analysis and scale reliability

Before calculating Cronbach's Alpha and conducting factor analysis, item characteristics were investigated individually for each of the seven cadaver images independently. Cronbach's Alpha was then calculated to evaluate the internal reliability of the refined scale after removing redundant items included within 24 scale items, Alpha coefficient ranged from 0.791 to 0.889.

Factor analysis

KMO values were calculated first, KMOs ranged from 0.710 to 0.838; this suggests good correlations. Principle component analysis demonstrated a single dominant factor for all 7 images which accounted for highest variance (%), ranging from 24 to 32.7. Similar to the phantom data, the highest variance factors explained the important information about the quality, meaning that these factors were highly loaded (≥ 0.4) into the majority of items of different themes (e.g. anatomical and others).

Rotated factor analysis demonstrates that factor numbers 1, 2, 3, and 4 for all the seven images accounted for approximately 12.5 ± 2 % of total variance within the scale (Table 6). These factors loaded high onto their items (≥ 0.4) and they are therefore considered significant factors to be included and support the scale validity (50, 51).

Table 6. An example of extracted factors from rotated factor analysis (image set 1).

Factor number	Total	% of Variance	Cumulative %
1	3.242	14.738	14.738
2	2.827	12.852	27.59
3	2.771	12.595	40.186
4	2.391	10.867	51.053
5	2.161	9.823	60.875

To improve factor analysis, scores of the 24 items for both phantom and cadaver were combined (i.e., all the completed scales (14 images x 335 volunteers = 2345)). Prior to this an unpaired t test was conducted to see if phantom and cadaver scores were significantly different; no significant difference existed ($P > 0.05$). Cronbach's Alpha for the combined phantom and cadaver images ranged from 0.786 to 0.907, KMO ranged from 0.734 to 0.908.

Discussion

Psychometric literature (10, 15, 27), supported by Bandura's guidelines (28), have provided a theoretical basis for the first ever *full development and validation* of a visual grading scale to assess perception of AP pelvis image quality. We propose that psychometric theory can be used to develop valid and reliable visual grading scales to assess image quality. This argument is supported by our scale having high reliability (Cronbach alpha ≥ 0.8), compact correlation patterns (KMO 0.7 to 0.9) and the ability to assess anatomical and technical quality, reflecting the construct validity. The argument is also supported by the excellent agreement between the visually and objectively predetermined quality rankings and the overall image quality scale scores (i.e. phantom: $r^2=0.94$, $p<0.001$; cadaver: $r^2=0.95$, $P<0.001$).

There is an overlap in the 95% CIs of the mean scale scores (see phantom images ranked 2 and 3, in Table 3; cadaver images 3 and 4, in Table 5). This suggests that scale sensitivity could be limited and it might not be able to differentiate between images of very similar quality. However, this kind of subtlety may not be a problem in real life situations and a slightly reduced scale sensitivity may not be critical (54, 55). Importantly, on inspection of these 4 images it is clear some have positional errors which the physical measure (i.e. SNR) would not have detected. Cases such as these demonstrate the limitation of simple physical measures (e.g. SNR) conducted in isolation of visual measures for the evaluation of image quality.

The final scale has 15 out of 24 anatomical items. These items relate to how clearly a given structure can be visualised in the image. The remaining items (9 out of 24) are positioning (e.g anatomical rotation) and technical (e.g. collimation errors). For the phantom images, removing these 9 items does not adversely affect the scale psychometric properties; the removal of scale items has similar outcomes within general psychometric literature (52, 56). This suggests that the AP pelvis scale has a uni-dimensional structure when used on phantoms. The explanation for this could be that the phantom images were uniform, aside having varying noise and contrast levels. On this basis it could be argued that phantoms can be used to create and validate visual grading scales. However, the use of phantoms in isolation could have limitations for

scale development and validation, as phantoms do not display a range of anticipated anatomical variations. If such variations are not taken into consideration then the scale could have limitations. For example, for some cadaver images Cronbach's alpha decreased when the 9 items were removed, thereby demonstrating their necessity. Consequently the use of human images should be considered in scale development and validation.

It is possible to derive subscales from our AP pelvis scale. Subscales are commonly derived from larger scales and published into the psychometric literature. For our AP pelvis scale an example of a subscale would be the 15 items highlighted in the appendix; these items can be used to assess AP pelvis anatomical image quality only. It is worth noting that subscales are not available with previously published visual grading scales in medical imaging, this highlights another novel value to the use of psychometric theory.

A normal approach in scale validation would be to assess a newly validated scale against an existing scale with known psychometric properties. Unfortunately no validated scale currently exists for AP pelvis image quality. Consequently additional work is warranted in order to further assess the validity of our scale.

Conclusion

No validated visual grading scale for AP pelvis image quality assessment exists within the literature; consequently many researchers rely on CEC criteria despite major deficiencies. In this paper, using AP pelvis as a catalyst, we have demonstrated that psychometric theory can be applied to develop and validate a visual grading scale for image quality. We propose that our method should be considered for developing and validating visual grading scales for other imaging procedures. Our 24 item scale shows a high level of internal reliability and validity. On this basis we propose our scale is likely to have value in clinical, research and education.

Appendix

AP pelvis image quality scale

- 1- The left hip joint is adequately visualised.
- 2- The right hip joint is adequately visualised.
- 3- The right lesser trochanter is visualized adequately.
- 4- The left lesser trochanter is visualized adequately.
- 5- The left greater trochanter is visualized adequately.
- 6- The right greater trochanter is visualized adequately.
- 7- The left iliac crest is adequately visualized.
- 8- The right iliac crest is adequately visualized.
- 9- The pubic and ischial rami are not adequately visualized.
- 10- The proximal femora are demonstrated adequately.
- 11- The left femoral neck is visualized adequately.
- 12- The right femoral neck is visualized adequately.
- 13- The left sacro-iliac joint is visualized adequately.
- 14- The right sacro-iliac joint is visualized adequately.
- 15- The sacrum and its intervertebral foramina are not visualized adequately.

Anatomical related items

- 16- There is appropriate differentiation between soft tissues.
- 17- The exposure factors used for this image are correct.
- 18- This image is sufficient for diagnostic purposes.
- 19- The medulla and cortex of the pelvis are adequately demonstrated.
- 20- The body of L5 is adequately demonstrated.
- 21- The obturator foramina are symmetrical.
- 22- Both acetabula are visualised clearly
- 23- The levels of rotation and axial tilting are within acceptable limits.
- 24- Fine bony detail is sufficiently demonstrated.

**Technical and
positioning related
items**

References

- (1) Båth M. Evaluating imaging systems: Practical applications. *Radiation protection dosimetry*. 2010; 139(1-4):26–36.
- (2) Lehnert T, Naguib N, Korkusuz H, Bauer R, Kerl J, Mack M, et al. Image-Quality Perception as a Function of Dose in Digital Radiography. *American Journal of Roentgenology*. 2011;197(6):1399-403.
- (3) Zarb F, Rainford L, McEntee M. Image quality assessment tools for optimization of CT images. *Radiography*. 2010;16(2):147-53.
- (4) Alsleem H, Davidson R. Quality parameters and assessment methods of digital radiography images. *The radiographer*. 2012;59(2):46-55.
- (5) Dowsett, D. J., Kenny, P. A., & Johnston, D. A. (2006). *The Physics of Diagnostic Imaging*. (2 ed.). Boca Raton: CRC Press.
- (6) Kundel, H.L. (1979). Images, Image Quality and Observer Performance: New horizons in radiology lecture. *Radiology*, 132(2), 265-271.
- (7) Seeram E, Bushong S, Davidson R, Swan H. Image Quality Assessment Tools for Radiation Dose Optimization in Digital Radiography: An Overview. *Radiologic technology*. 2014;85(5):555-62.
- (8) Li Y, Poulos A, McLean D, Rickard M. A review of methods of clinical image quality evaluation in mammography. *European Journal of Radiology*. 2010;74(3):122-31.
- (9) H. S, Moores BM. Diagnostic radiology better images - lower dose. Compromise or correlation? A European strategy with historical overview. *Journal Belge de Radiologie*. 1995;78(6):382-7.

- (10) Mraity H, England A, Akhtar I, Aslam A, De Lange R, Momoniat H, et al. Development and validation of a psychometric scale for assessing PA chest image quality: A pilot study. *Radiography*. 2014;20(4):312–7.
- (11) Sund P, Båth M, Kheddache S, Månsson LG. Comparison of visual grading analysis and determination of detective quantum efficiency for evaluating system performance in digital chest radiography. *European radiology*. 2004;14(1):48-58.
- (12) CEC. *European guidelines on quality criteria for diagnostic radiographic images*: (EUR 16260 EN). Brussels: CEC, 1996.
- (13) Honey I, Hogg P. Balancing radiation dose and image quality in diagnostic imaging. *Radiography*. 2012;18(1):1-2.
- (14) ICRP. Patient dose and image quality in digital radiology: ICRP Publication 93. *Annals of ICRP*. 2004;34(1):21-49.
- (15) Mraity H, England A, Hogg P. Developing and validating a psychometric scale for image quality assessment. *Radiography*. 2014;20(4):306–11.
- (16) Busch HP, Jaschke W. Adaptation of the quality criteria concept to digital radiology. *Radiation protection dosimetry*. 1998;80(1-3):61-3.
- (17) Moores BM, Mattsson S, Månsson LG, Panzer W. Quality criteria development within the fourth framework research programme. *Radiation protection dosimetry*. 2000;90(1-2):63–71.
- (18) Coaley K. *An introduction to psychological assessment and psychometrics*. Los Angeles: Sage Publications; 2010.
- (19) Chan CTP, Fung KLL. Dose optimization in pelvic radiography by air gap method on CR and DR systems e A phantom study. *Radiography*. 2015;21(3):214–23.

- (20) Sherer MAS, Visconti PJ, Ritenour ER, Haynes KW. Radiation protection seventh edition in medical radiography. 7 ed. Philadelphia: Mosby; 2014.
- (21) Hart D, Wall BF, Hillier MC, Shrimpton PC. Frequency and collective dose for medical and dental X-ray examinations in the UK. Chilton: Health Protection Agency 2010 Contract No: HPA-CRCE-012.
- (22) Al Khalifah A, Brindhavan A. Comparison between conventional radiography and digital radiography for various kVp and mAs settings using a pelvic phantom. *Radiography*. 2004;10(2):119–23.
- (23) England A, Evans P, Harding L, Taylor EM, Charnock P, Williams G. Increasing Source-to-Image Distance to Reduce Radiation Dose From Digital Radiography Pelvic Examinations. *Radiologic technology*. 2015;86(3):246-56.
- (24) Persliden J, Beckman K, Geijer H, Andersson T. Dose-image optimisation in digital radiology with a direct digital detector: an example applied to pelvic examinations. *European Radiology Supplements*. 2002;12(1):1584-8.
- (25) Heath R, England A, Ward A, Charnock P, Ward M, Evans P, et al. Digital Pelvic Radiography: Increasing Distance to Reduce Dose. *Radiologic technology*. 2011;83(1):20-8.
- (26) Manning-Stanley AS, Ward AJ, England A. Options for radiation dose optimisation in pelvic digital radiography: A phantom study. *Radiography*. 2012;18(4):256-63.
- (27) Spector PC. *Summated rating scale construction* : An introduction. California: Sage Publications; 1992.
- (28) Bandura A. *Self-efficacy beliefs of adolescents*. Pajares F, Urdan T, editors. New York: Information Age; 2006.

- (29) Abell N, Springer DW, Kamata AO. Developing and validating rapidassessment instruments. Tripodi T, Dattalo P, Thyer BA, Danto EA, Harrington A, Webster JM, et al., editors. Oxford Oxford University Press; 2009.
- (30) Smolleck LD. *The development and validation of an instrument to measureperspective teachers's self-efficacy in regard to the teaching of science as inquiry*. PhD thesis. USA: The Pennsylvania State University; 2004.
- (31) Field A. Research methods II: project 2, questionnaire design [Lecture note];2005.
- (32) Fauber TL. *Radiographic imaging & exposure*. 4 ed. Missouri: Mosby; 2014.
- (33) Mansson LG. Methods for the evaluation of image quality. *Radiation protection dosimetry*. 2000;90(1-2):89–99.
- (34) RSNA. RadLex: a lexicon for uniform indexing and retrieval of radiology informationresources 2010 [cited 2013 January 10]. Available from: <http://www.rsna.org/radlex/>.
- (35) ILO. Guidelines for the use of the ILO international classification of radiographsof pneumoconioses. Revised Edition 2000. In: Occupational safety and healthseries, No. 22. Geneva: International Labour Office; 2002.
- (36) Tapiovaara M. Relationships between physical measurements and user evaluation of image quality in medical radiology - a review. STUK-A219. 2006:1-2.
- (37) Whitley, A. S., Sloane, C. , Hoadley, G., Moore, A. D., & Alsop, C. W. (2005). *Clark's positioning in radiography* (12 ed.). London: Hodder Arnold.
- (38) Hiles P, Mackenzie A, Scally A, Wall B. Recommended standards for the routine performance testing of diagnostic X-ray imaging systems Institute of Physics and Engineering in Medicine ;Report NO 91. 2005.

- (39) Brennan PC, McEntee M, Evanoff M, Phillips P, O'Connor WT, Manning DJ. Ambient lighting: effect of illumination on soft-copy viewing of radiographs of the wrist. *American roentgen ray society*. 2007;188(2):177-80.
- (40) Brenowitz N, Tuttle C. Development and testing of a nutrition-teaching self-efficacy scale for elementary school teachers. *Journal of nutrition education and behavior*. 2003;35(6):308-11.
- (41) Tabachnick B, Fidell L. Using multivariate statistics. 5 ed. Boston: Pearson Education; 2013.
- (42) Ware J, Gandek B. Methods for Testing Data Quality, Scaling Assumptions, and Reliability: The IQOLA Project Approach. *Journal of clinical epidemiology*. 1998;51(11):945-52.
- (43) Field A. *Discovering statistics using SPSS for windows*. London: Sage Publications; 2000.
- (44) Nunnally J, Bernstein I. *Psychometric theory*. 3 ed. New York: McGRAW-HILL; 1994.
- (45) Cronbach L. *Coefficient alpha and the internal structure of tests*. *Sychometrika*. 1951;16 (1):297-334.
- (46) Streiner DL, Norman GR. *Health measurement scales: a practical guide to their development and use*. 4 ed. USA: Oxford University Press; 2008.
- (47) Koutra K, Orfanos P, Roumeliotaki T, Kritsotakis G, Kokkevi A, Philalithis A. *Psychometric Validation of the Youth Social Capital Scale in Greece*. *Research on social work practice*. 2012;22(3):333-43.
- (48) Filliben J. Exploratory Data Analysis. C., Croarkin & W., Guthrie, NIST/SEMATECH e-Handbook of Statistical Methods (Web-based book) . Retrieved

from <http://www.itl.nist.gov/div898/handbook/eda/section3/eda35h1.htm>.
2012.

(49) Kaiser SF. An index of factorial simplicity. *Sychometrika* 1974;39(1):31-6.

(50) Field A. Research methods II:Project 2.Factor analysis on SPSS. [Lecture notes];2005.

(51) Suhr DD, editor Exploratory or Confirmatory Factor Analysis? SUGI 31 Proceedings; 2006; San Francisco, California: SAS Institute Inc; 2006.

(52) Kitching J, Cassidy S, Hogg P. Creating and Validating Self-Efficacy Scales for Students. *Radiologic technology*. 2011;43(1):10-9.

(53) Dennis CL, Faux S. Development and Psychometric Testing of the Breastfeeding Self-Efficacy Scale. *Research in Nursing & Health*. 1999;22(5):399–409.

(54) Grahn A, Hemdal B, Andersson I, Ruschin M, Thilander-Klang A, Börjesson S, et al. Clinical evaluation of a new set of image quality criteria for mammography. *Radiation protection dosimetry*. 2005;114(1-3):389–94.

(55) Joyce M, McEntee M, Brennan P, O’Leary D. Reducing Dose for Digital Cranial Radiography: The Increased Source to the Image-receptor Distance Approach. *Journal of Medical Imaging and Radiation Sciences*. 2013;44(4):180–7.

(56) Cassidy S, Eachus P. Developing the computer user self-efficacy (cuse) scale: Investigating the relationship between computer self-efficacy, gender and experience with computers. *Journal of educational computing research*. 2002;26(2):133-53.