Title: Evaluate the reliability of the apprenticeship in the first year of medical school: towards a reliable first level ultrasound examination.

Michela Cristina Turci¶¹, Massimo Tassoni¶², Florida Gajtani&², Antonino Michele Preverita&³, Alberto Magenta-Biasina&⁴, Elia Mario Biganzoli¶⁵ and Graziano Serrao*¶²,³.

¹Dept. of Biomedical Sciences for Health. Università degli Studi di Milano. Via Mangiagalli, 31 I-20133 Milan, Italy; ²SPUMS San Paolo Ultrasound Medical School, San Paolo Hospital. Università degli Studi di Milano. Via Antonio di Rudini, 8 I-20142 Milan, Italy; ³Dept. of Health Sciences. Università degli Studi di Milano. Via Antonio di Rudini, 8 I-20142 Milan, Italy; ⁴Servizio di Radiologia Diagnostica e Interventistica. Azienda Ospedaliera Santi Paolo e Carlo, Ospedale San Paolo. Via Antonio di Rudini, 8 I-20142 Milan, Italy; ⁵Dept. of Clinical Sciences and Community Health. Università degli Studi di Milano. Via Augusto Vanzetti, 5 I-20133 Milan, Italy.

* Corresponding author

E-mail: graziano.serrao@unimi.it; iD: 0000-0001-9651-7685

¶ These authors contributed equally to this work

& These authors also contributed equally to this work

Conflict of interests: The authors declare that they have no proprietary, commercial, or financial interests that could be construed to have inappropriately influenced this study.

Funding: NONE

Authors Contributions

Conceptualization: Graziano Serrao

Data Curation: Massimo Tassoni, Michela Cristina Turci

Investigation: Massimo Tassoni, Florida Gajtani, Alberto Magenta Biasina, Antonino Michele Preverita

Formal Analysis: Elia Mario Biganzoli
Writing – Original Draft Preparation Michela Cristina Turci, Graziano Serrao

Writing – Review & Editing: all authors
Abstract

Objectives
Our aim was to motivate apprentices’ sonographer needs, to appraise their own measurements, to reduce inconsistencies within and between operators. Deep knowledge of ultrasound sectional anatomy is mandatory for an appropriate performance.

Methods
In three different weekdays, 3 sonographer apprentices (rater), randomly selected from a cohort of San Paolo Medical School first year students participated in vertically integrated study of living anatomy through ultrasound examination, repeated lumbar multifidus cross-sections scans on 6 subjects at lumbar level. The Agreement R package 0.8-1 was used to monitored the performances of each apprentice.

Results
The agreement (CCC_{intra} 0.6749; CCC_{inter} 0.556; CCC_{total} is 0.5438) was further from least acceptable CCC of 0.92-0.95. The precision indices (precision_{intra} 0.6749; inter 0.801; total 0.6274) were unsatisfactory, while the accuracy was high (0.9889 to 0.9913). The same occurred for the agreement on rater performances comparisons, where readings were high accurate (0.9537 to 0.9733) but moderately precise (0.7927 to 0.8895), not interchangeable TIR (1.173) but without rater supremacy. IIR (r1 vs r2 1.104, r1 vs r3 1.015, r2 vs r3 0.92) 95% confidence limits.

Conclusions
Apprentices were not reliable, repeatable, interchangeable. The weak link in the method seemed to be cultural weakness on vivo imaging morphologies, qualitative and quantitative measurement procedure on elementary statistical processing.

**Introduction**

Moved by evidence, teachers must work hardly to transform excellence into routine performances. Handle echography represents a key toll that allows an immediate correlation between imaging findings and clinical data, which will improve the management of prime care supply at the patient’s home. The fast advancement on telemedicine, the increased use and the decreasing instrument costs lead to a great demand for appropriate education toward a sturdy echo’s competence [1-5]. Echo continuously improves the instrumental semiotics potentials by making more anatomic structures visible. Since 2007 at San Paolo Medical School, an ultrasound learning resource was located within the anatomy area [6]. The focus of this educational training would be the capacity to forge a physician able to execute a first level ultrasound examination (FLUX). Often in clinical practice, only one measurement per patient is performed, hence before getting down to the data analysis you want to ensure that they are not contaminated by assessed procedure/factors.

Our aim was to verify apprentice’s judgement validity and tune learning path.

**Methods**

Ultrasonic method is commonly described as largely employable, pliant, and non-invasive [6,7]. In 2009, San Paolo Medical School Didactic College planed vertically integrated study of the living body anatomy through peer physical and ultrasound examination. Core group of faculty experts in ultrasound developed a standardized curriculum that was presented to the student body. Faculty-supported facility for medical students' independent education projects allowed each student to be alternately examine and examiner [8]. To evaluate the competences acquired and tune the training on academic year 2019-2020, 3 women and 3
men (ranging in age 19 to 22±1.27yr, and in BMI 19.0 to 25.1±2.42 kg×m⁻²) healthy apprentices, ruffled among participants in the ultrasound living anatomy course took part in the present study as examines, but only 3 as examiners too. The Ethics Committee of the University of Milan examined and issued, an ex post favourable opinion on the project (03/28/2022 advisory 31/22).

**Scanning procedure**

An experienced physiatrist (AMP) evaluated each drawn student to exclude any sequels that could alter the region of interest. A multi-frequency probe equipped ultrasound machine B-mode Real Time (Logic QE, Medical Systems, Milwaukee, WI, USA) was used. A convex probe at 5 MHz emission was employed [9,10], 3 apprentices (raters), under conditions as close as possible to the clinical daily routine, performed ultrasound scans then run measurement/rating process.

Following the procedure commonly recommended in literature, each rater proceeded by steps. To flatten down the lumbar lordosis the subject lied prone with a pillow under the pubic bones and feet off the bed, arms along her/his hips and head turned toward the most comfortable side (Fig. 1). To identify the spinous processes of L4-L5 the rater aligned each of her/his index fingers with the correspondent iliac crest and extended the thumbs towards the spinous processes of the vertebral column (Fig. 2) [11]. To confirm the level found by palpation, each rater verifies by ultrasound scan [12]. In order to allow the vertebral lamina, the multifidus muscle was scanned longitudinally (Fig. 3).

Figure 1. Subject position.

Figure 2. Identification of the spinous processes.

Figure 3. Lumbar longitudinal scan between L4 and sacrum a: apex of the spinous process, b: yellow ligament, c: median sacral crest.
Thereafter in order to be transverse correctly aligned, the rater performed a 90° rotation of the probe. The discrimination of lateral edges of multifidus out of the surrounding muscles was the most difficult part of the scanning protocol. Each subject was requested of slightly lifting the ipsilateral lower limb from the couch; the identification of that lateral edge was so favored. After the contraction, before measuring the muscle diameters, the rater waited a few seconds to settle the rest position [13,14]. After having identified the multifidus, the rater moved the probe to the left and then to the right while tracing antero-posterior (r/lAP) and latero-lateral diameter (r/lLL) assessments. On frozen images, the AP diameter, from the lamina to the inner edge of the upper edge of the multifidus, and the LL diameter, corresponding to the maximum width from the lateral edge of the multifidus to the spine, were marked out. Each rater established independently the positioning of the subject and the anatomical landmarks. Raters were not aware to other rater’s judgments. In accord with Watson statements [15,16], 3 repeated measurements are usually considered appropriate, hence on different weekdays, the procedure was 3 times repeated, on 6 subjects, at five vertebral levels, on each side and therefore 180 scans generated 360 diameters.

**Statistical analysis**

Reliable and accurate measurements have a common theme: assessing closeness (agreement) between observations. Because there are 3 raters involved, one may like assessing the intra, inter and the total agreement with replicated readings produced by different observers. A meaningful statistic to measure the agreement of observations has been the mean squared deviation (MSD). The *Concordance Correlation Coefficient* (CCC), a scaled index proposed by Lin [17,18], translates the MSD into a combine’s coefficients of precision (degree of refinement) and accuracy (degree of total displacement) that can be used to compare the differences between operators and sessions. To esteem multiple observers with replications, where none of the observers is treated as reference, Barnhart [19] proposed $\text{CCC}_{\text{total}}$, $\text{CCC}_{\text{inter}}$, $$...$$
and CCC\textsubscript{intra} series of indices. Another intuitive agreement index is a measure that captures a large proportion of data within a settled maximum acceptable absolute difference between two observers’ readings. This probability is called coverage probability (CP). On the other hand, if we set a predetermined coverage probability, we can find the boundary so that the probability of absolute difference is less than the boundary itself. This boundary called total deviation index (TDI) is proportional to the square root of the MSD and it is a performance measurement. A satisfactory agreement may require a large CP or equivalently a small TDI [18,20,21].

When multiple raters are available with replicates, we are often interested if raters are interchangeable. The Total-Intra Ratio (TIR) assesses comparative agreement indices for the interchangeability between multiple raters with multiple readings. The scale criterion means that the MSD\textsubscript{total-raters} cannot be more than a predefined value of the MSD\textsubscript{intra-rater}, or TIR\textsubscript{R} < predefined value, with 95% confidence. In addition, we compared the intra rater precision by computing MSD\textsubscript{intra1_2}/MSD\textsubscript{intra3}, MSD\textsubscript{intra1_3}/MSD\textsubscript{intra2}, and MSD\textsubscript{intra2_3}/MSD\textsubscript{intra1} or Intra-Intra Ratio (IIR). A 100(1 − \alpha/2) % confidence interval for IIR is computed based on the log-transformed IIR estimates and claim the superiority or inferiority if the upper or lower limit is less than or greater than 1.0 [22]. All index was computed by Agreement R package 0.8 [23].

Results

Table 1 shows the multifidus cross section AP and LL diameters measurements (mm) relieved by each rater, for each subject and lumbar level.
Based on multifidus assessments conducted by experts in the field found out on literature in terms of agreement, we set a within-rater deviation not more than 30% of the total deviation, and inter-rater deviation not more than 50% of the total deviation. These translated into a least acceptable CCC_intra of 0.91 (1 – 0.3²), and a least acceptable CCC_inter of 0.75 (1 – 0.5²). 
The CCC estimates are shown on Table 2.

Table 2. CCC estimates (see text for details).

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Estimate</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCC&lt;sub&gt;intra&lt;/sub&gt;</td>
<td>0.674</td>
<td>0.597</td>
</tr>
<tr>
<td>CCC&lt;sub&gt;inter&lt;/sub&gt;</td>
<td>0.792</td>
<td>0.732</td>
</tr>
<tr>
<td>CCC&lt;sub&gt;total&lt;/sub&gt;</td>
<td>0.621</td>
<td>0.627</td>
</tr>
<tr>
<td>Precision&lt;sub&gt;intra&lt;/sub&gt;</td>
<td>0.674</td>
<td>0.597</td>
</tr>
<tr>
<td>Precision&lt;sub&gt;inter&lt;/sub&gt;</td>
<td>0.801</td>
<td>0.744</td>
</tr>
<tr>
<td>Precision&lt;sub&gt;total&lt;/sub&gt;</td>
<td>0.627</td>
<td>0.550</td>
</tr>
</tbody>
</table>

The CCC<sub>intra</sub> was estimated to be 0.6749, which means a within-sample deviation is about 57% $\sqrt{(1 - 0.6749)}$ of the total deviations. The CCC<sub>inter</sub> was estimated to be 0.7921, which means a within-sample deviation is about $\sqrt{(1 - 0.791)}$ 63% of the total deviations [17]. The CCC<sub>total</sub> was estimated to be 0.6219, which means for individual observations from different rater, the within-sample deviation is about 61% of the total deviations. These findings are consistent with moderate precision (0.6749<sub>intra</sub>, 0.8<sub>inter</sub>, 0.6274<sub>total</sub>), but high accuracy (0.9889<sub>inter</sub> to 0.9913<sub>total</sub>) ever larger than 0.95 and 0.91, respectively. Performance analysis of raters is shown in Table 4.

Table 3. Rater performances (see text for details).

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Estimate</th>
<th>95% CI</th>
<th>Allowance</th>
<th>RBS</th>
</tr>
</thead>
<tbody>
<tr>
<td>TDI&lt;sub&gt;95 r1-2&lt;/sub&gt;</td>
<td>5.573</td>
<td>6.19</td>
<td>5.4</td>
<td></td>
</tr>
<tr>
<td>TDI&lt;sub&gt;95 r1-3&lt;/sub&gt;</td>
<td>8.494</td>
<td>9.453</td>
<td>5.4</td>
<td></td>
</tr>
<tr>
<td>TDI&lt;sub&gt;95 r2-3&lt;/sub&gt;</td>
<td>8.608</td>
<td>9.581</td>
<td>5.4</td>
<td></td>
</tr>
<tr>
<td>CP&lt;sub&gt;54 r1-2&lt;/sub&gt;</td>
<td>0.941</td>
<td>0.908</td>
<td>0.95</td>
<td>0.234</td>
</tr>
<tr>
<td>CP&lt;sub&gt;54 r1-3&lt;/sub&gt;</td>
<td>0.783</td>
<td>0.730</td>
<td>0.95</td>
<td>0.03</td>
</tr>
<tr>
<td>CP&lt;sub&gt;54 r2-3&lt;/sub&gt;</td>
<td>0.777</td>
<td>0.723</td>
<td>0.95</td>
<td>0.01</td>
</tr>
<tr>
<td>TIR&lt;sub&gt;r1-2-3 vs all&lt;/sub&gt;</td>
<td>1.173</td>
<td>1.72</td>
<td>1.5</td>
<td></td>
</tr>
<tr>
<td>IIR&lt;sub&gt;r1 vs r2&lt;/sub&gt;</td>
<td>1.104</td>
<td>(0.842,1.447)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>IIR&lt;sub&gt;r1 vs r3&lt;/sub&gt;</td>
<td>1.015</td>
<td>(0.687,1.499)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IIR&lt;sub&gt;r2 vs r3&lt;/sub&gt;</td>
<td>0.92</td>
<td>(0.633,1.336)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

RBS relative bias squared.

The agreement was moderate and further from least acceptable CCC of 0.91-0.95. This finding agrees with the data presented in Fig. 4a-c, where readings were evenly scattered around the 45°
line (high accuracy 0.9537 to 0.9733) but not tightly scattered (moderate precision 0.7927 to 0.8895). The plots indicated that score differences tended to be more dispersed for large value correspondent to a lower lumbar level.

**Fig 4. Inter-rater (R) agreement: a R1vsR2; b R1vsR3; c R2vsR3.**

In terms of TDI and CP indices, the least acceptable agreement was set as having at least 95% pair’s observations within 5.4 mm (allowance, equal to 20% of all measurements mean), independently to rater. The TDI$_{0.95r1r2}$ estimate was 5.573 mm, close to 5.4 mm target values, which means based on the average readings, 95% of the readings are within 5.573 mm of their replicate readings from the other rater. The TDI$_{0.95r1r3}$ estimate was 8.494 mm, and the TDI$_{0.95r2r3}$ estimate was 8.608 mm. The one-sided upper confidence limit for TDI$_{0.95r1r2}$ was 6.19 mm, TDI$_{0.95r1r3}$ 9.453 mm, and TDI$_{0.95r2r3}$ 9.581 mm all larger than target values. The CP$_{54r1r2}$ estimate was 0.9418, which means that 94% of observations were within their target values from the other rater. The one-sided lower confidence limit for CP$_{54r1r2}$ was 0.9170, smaller than 0.95. The CP$_{54r1r3}$ estimate was 0.7835 and CP$_{54r2r3}$ 0.7773 which means that 78% of observations are within 5.4 mm. The one-sided lower confidence limit for CP$_{54r1r3}$ was 0.7301, and for CP$_{54r2r3}$ was 0.7236 both smaller than 0.95.

TIR$_{r1-2-3}$ vs all the ratio between average of the total MSD of all raters relative to the average of intra MSD of all raters, was estimated to be 1.173, with a 95% upper confidence limit of 1.72. If we consider that a precision deviation of more than 50% (1.5 allowance) is clinically not acceptable, we cannot claim raters’ interchangeability. The IIR was estimated to be IIR$_{r1vsr2}$ 1.04, IIR$_{r1vsr3}$ 1.015 and IIR$_{r2vsr3}$ 0.92. If we refer at the individual bioequivalence FDA criterion, that the ratio of geometric means between readings differences must lie between 0.8 and 1.25, IIR upper/lower 95% interval {r1 vs r2 (0.842, 1.447), r1 vs r3 (0.687, 1.499), r2 vs r3 (0.633, 1.336)} indicated no rater precision supremacy.
Discussion

The study confirmed the ultrasound procedure as an operator dependent technique that is prone to variable measurements. Each considered parameter (rater and replicates) induced bias. Each rater proved to be accurate but not sufficiently precise.

Lacking intra-operator and inter-operator reliability made impossible to compare data obtained by different raters of the same department or in different groups and centers, and also frustrated a complete tuning of correct guidelines. Because errors are inherent in every measurement procedure, one must ensure the magnitude of the measurement agreement [18,24,25]. In clinical practice measurements are detect to take a decision, consequently magnitude of acceptable differences between ratings, it depends on clinical decision consequences. No universal standard for an acceptable confidence interval is currently available, but most experts consider range values 0.8-1.25 as supporting good reliability. A confidence interval for the difference between measurement results indicates the smallest detectable difference (SDD) [26-29]. With repeated rating, any change outside these boundaries can be considered a true change in the entity being assessed. The smallest detectable difference it is a useful starting point in determining what is a minimum clinically important difference (MCID) in values handle.

When measurements show evidence of lack of agreement, we need to address the sources of the deficiencies. According to Alvan Feinstein [30] for gaining insight into the “clinimetric” property, we should reinforce apprentice’s attention on measurements methodologic discipline quality of the measurement [31,32]. To enable assessment in routine care, we’ll beckon worth on agreement, reliability, repeatability, and reproducibility. Because of that, it is essential to reinforce statistical skill in medical training too.

Conclusions

The cohort sampled was limited but indicative to same training adjustments. Future efforts could focus on tear down “observational errors”. In order to improve rater’s performance, and promote inter-rater reliability, it should be considers a number of strategies: training on imaging
morphologies, nurture feedback to those who have shown a low reproducibility, reinforce protocols. All these methods used as a screening quality, self-evaluation check, and calibration experience encourage raters to achieve 0.8-1.25 range values and gaining smallest detectable difference validity [27,33].
References


https://doi.org/10.1371/journal.pone.0185031


34. Figures

35. Figure 4. Subject position.

36. Figure 5. Identification of the spinous processes.

37. Figure 6. Lumbar longitudinal scan between L4 and sacrum

40. a: apex of the spinous process, b: yellow ligament, c: median sacral crest.

46. Fig 4. Inter-rater (R) agreement: a R1 vs R2; b R1 vs R3; c R2 vs