Training using a simulation-based workshop reduces inaccuracies in estimations of testicular volume

CRAIG, Jessica, SHARMAN, Megan, FITZGERALD, Ciara, WIGG, Dominic, BETH, Williams, WILKINSON, Ellen, WRIGHT, Neil, LANGLEY, Joseph <http://orcid.org/0000-0002-9770-8720> and ELDER, Charlotte

Jessica N. Craig, Megan R. Sharman, Ciara G. Fitzgerald, Dominic Wigg, Beth S. Williams, Ellen E. Wilkinson, Neil P. Wright, Joe Langley and Charlotte J. Elder*

Training using a simulation-based workshop reduces inaccuracies in estimations of testicular volume

https://doi.org/10.1515/jpem-2020-0312
Received May 28, 2020; accepted September 21, 2020;
published online November 11, 2020

Abstract

Objectives: Measuring testicular volume (TV) by orchidometer is routine in the clinic when staging male puberty. We have developed a simulation model for TV estimation and investigated whether training medical students, using a workshop with simulation models, could improve the accuracy and reliability of TV estimation.

Methods: All participating medical students watched a video representing standard undergraduate training in male pubertal assessment. Volunteers were then randomised directly to assessment or to attend a workshop consisting of a further video and four stations contextualising and practising the skills required for TV estimation, prior to assessment. Three child mannequins displaying testes of 3 mL, 4 mL (twice), 5, 10 and 20 mL were used for assessment. Participants were asked to return a fortnight later for repeat assessment to assess intra-observer reliability, the effect of repeated examinations on accuracy and time on skill retention.

Results: Ninety students participated (55F), 46 attended the workshop and were considered "trained". There was no difference between the groups in numbers of correct estimations (29% trained, 27% untrained, p=0.593). However, the trained group's estimations were closer to the true volume, with more from the trained group one bead away (p=0.002) and fewer more than three beads away from the true volume (p<0.001), compared to the untrained group. Trained participants were more accurate at the second assessment (n=80) (p<0.001) and had greater intra-observer reliability (p=0.004).

Conclusions: Overall TV estimation accuracy was poor. Workshop-style training improved accuracy, reliability and retention of skill acquisition and could be considered as a useful learning tool.

Keywords: accuracy; simulation; pubertal stage; reliability; testicular volume; training.

Introduction

Testicular volume (TV) is an essential clinic tool in the staging of male pubertal development. TV is also assessed by urologists and andrologists as a surrogate marker of testicular function, since 80–90% of testicular mass comprises of seminiferous tubules, the site of spermatogenesis [1]. Calipers and ultrasound can be used to assess TV, but the commonest method is the Prader orchidometer, consisting of a string with ellipsoid beads of increasing sizes from 1–25 mLs, which are used to compare against the patient's testis [1, 2]. Although ultrasound is a more precise method in the assessment of TV, the use of a Prader orchidometer is more practical and less expensive, additionally providing clear decision cut points in the assessment of precocious puberty [2]. The onset of male puberty is defined as a TV of 4 mLs, peak height velocity is seen at TVs of 12–15 mLs and adult TVs can range between 12 and 25+ mLs [3]. Accuracy of TV estimation, particularly at the lower volumes, is important. Missing the diagnosis of male precocious puberty (onset before the age of nine years) may have serious sequelae, due to the likelihood of pathological aetiologies including brain tumours [4].

Despite its popularity and importance, it is acknowledged that the accuracy and reliability of TV estimation using a Prader orchidometer may be poor [1, 2, 5]. A recent

Joe Langley and Charlotte J Elder are joint senior authors.

***Corresponding author: Dr. Charlotte J. Elder**, Senior Lecturer in Paediatric Endocrinology Academic Unit of Child Health, Sheffield Children's NHS Foundation Trust, Sheffield, UK, Phone: +44 (0) 114 2260716, E-mail: c.j.elder@sheffield.ac.uk. https://orcid.org/0000-0003-2390-5593

Jessica N. Craig, Megan R. Sharman, Ciara G. Fitzgerald, Beth S. Williams and Ellen E. Wilkinson, The University of Sheffield, Sheffield, UK

Dominic Wigg and Joe Langley, Sheffield Hallam University, Sheffield, UK

Neil P. Wright, The University of Sheffield, Sheffield, UK; and Sheffield Children's Hospital, Sheffield, UK

study found that significantly larger volumes were obtained using the Prader orchidometer than with ultrasound and intra-operative caliper measurements [6]. We have previously reported the largest study to date assessing the accuracy and reliability (inter- and intra-observer) of TV estimation in clinicians and the only study to use simulation as the assessment tool. In collaboration with design engineers we developed a simulation model for TV estimation with pre-, peri- and post-pubertal sized silicon testes, housed in latex scrotum and displayed on paediatric mannikins. We tested 215 delegates at a meeting of Paediatric Endocrinologists using six testes and five different volumes. TV was estimated correctly on only 33% of occasions and intra-observer reliability was lacking with individuals giving different estimations for the same size testicle on 61% of occasions. Experience of clinicians was found to improve accuracy. Clinicians had a tendency to confer biological symmetry, by overestimating the 4 mL testis when paired with a 5 mL and underestimating it when paired with a 3 mL testis [5].

The use of simulation to teach healthcare professionals basic and advanced skills, such as resuscitation and airway management, is becoming increasingly sophisticated and mainstream as an effective education method [7, 8]. Retention of information and its application are strongest when taught and practised in situations which are of a similar setting to real life and in the workplace [8]. We describe a study to train endocrinology-naïve medical students, using our simulation models in a workshop format, to assess whether improvement in the accuracy and reliability of TV estimation can be demonstrated.

Materials and methods

Study population and recruitment

We invited medical students from the University of Sheffield, UK, to participate in the study. We chose medical students as they are naïve to TV estimation and could therefore act as surrogates for new trainees in paediatric endocrinology, whilst providing adequate numbers required for the study. We used the data from our previous pilot study to inform our power calculation. The accuracies of trainees attending a puberty teaching session, and considered the "untrained" population (12% correct estimations), and consultant paediatric endocrinologists, considered the "trained" population (39% correct estimations), were used to calculate that 76 students would be required for the study ($p<0.05$) [5, 9]. The study participants were recruited on a voluntary basis through advertisement on the University intranet and at student lectures. As no one attending the sessions was excluded, we exceeded the study's recruitment target. Ethical approval was not required for the study as it recruited students in their professional and learning role and did not involve patients.

Workshop design

We invited eligible students to attend one of six sessions, with a maximum of 30 students per session. At the start of each session, all participants watched a 5-min video recording of the undergraduate teaching delivered to third-year medical students on male pubertal staging and the Prader orchidometer, given by a senior consultant paediatric endocrinologist (NPW). This represented the knowledge level expected of a newly qualified doctor. We then divided the cohort, using a random number generator (https://www.random.org/lists/), to create two groups with equal sex distribution. One group proceeded to the workshop to become "trained", and the other went directly to the assessment "untrained".

Although targeting multiple learning styles to maximise the impact of teaching is widely advocated, this approach lacks a robust scientific evidence base [10, 11]. However, students often have a preference for how they perceive they learn best and, informed by the feedback from our pilot study, we designed the workshop around these different learning preferences [11, 12]. Four different learning modalities were used: visual (instructions, video demonstrations), auditory (instructions, video demonstration, participant discussion), tactile (feeling workshop items and completing worksheets) and kinaesthetic (active involvement, moving around workshop stations) [12]. In addition, we encouraged students to reflect upon their experience through active discussion with each other, a recommendation for embedding learning [13]. To control for variation between the different groups' sessions and ensure reproducibility, we ran each workshop using an identical format, with scripted instructions and defined, non-transferable, roles for the three session moderators.

At the start of the workshop, participants watched another 2:30 min video recording (NPW) contextualising the examination and its importance, then explaining how to correctly use the Prader orchidometer. They then progressed through four workshop stations, each lasting 5 min, in a pre-specified order to take them from abstract testicular metaphors to "real-life" simulation models. At station 1, participants were given nine food substances, including olives, cocktail sausages and different sized baby potatoes, as real-life metaphors to relate to orchidometer bead volumes (Figure 1A). Station 2 introduced the orchidometer, using it to estimate five prosthetic testes of different volumes (Figure 1B). The testes were not constrained by a scrotum to focus participants on volume and shape. At station 3, there were 12 sets of varying plastic testes in fabric bags to introduce the effect of the scrotum on TV estimation and remove the visual aid, working towards a kinaesthetic skill (Figure 1C). Station 4 involved participants measuring the TV of six simulation testes attached to mannikins, similar to those being used in the assessment (Figure 1D). Throughout the workshop, to enhance learning, we provided students with immediate feedback through flip sheets displaying the correct TVs. This technique has been shown to improve students' learning experience and enhance educational attainment [14]. We encouraged participants to record their answers in a worksheet to aid visual/non-verbal learners and to further embed knowledge. Students reflected upon their active experience through discussion of their findings with fellow participants (peer learning) and questions from the study team pertaining to over- or underestimation and the realism of the learning materials.

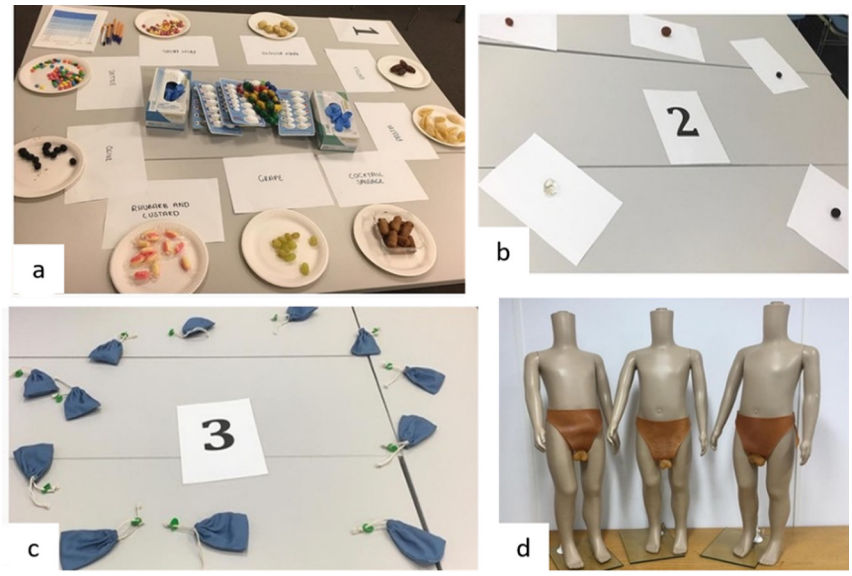

Figure 1: Simulation training workshop (Figure 1A Workshop station 1: food substances as real-life metaphors to relate to orchidometer bead volumes. Figure 1B Workshop station 2: prosthetic testes. Figure 1C Workshop section 3: 12 sets of varying sized plastic testes in fabric bags. Figure 1D Simulation models displayed on paediatric mannikins for TV estimation used in workshop section 4 and assessment.)

Assessments

Participants recorded their estimates together with their basic demographics (age, sex and medical school year) and whether they had previously received any TV estimation training.

The assessment consisted of three child mannequins adorned in latex briefs with latex scrotum containing specifically engineered silicon testes (Figure 1D). We have previously published details about the development and validation of these materials [5].

- Mannequin 1 – Left (4 mL) Right (5 mL)
- Mannequin 2 – Left (3 mL) Right (4 mL)
- Mannequin 3 – Left (10 mL) Right (20 mL)

We chose testicular sizes to reflect clinically important thresholds. We selected 4 mL twice to look at the effect of estimation when paired with a smaller and larger testis.

To assess intra-observer reliability, the effect of repeated examinations on accuracy and the effect of time elapsed on skill retention we asked participants to return a fortnight later for a repeat assessment. In order not to compromise the results of the second assessment, we did not inform participants of the true testicular volumes at any point during or after the first assessment.

Data analysis

We analysed data using SPSS for Mac Version 16.28 and Prism 8 for macOS version 8.4.1. The difference between participants' estimates and the TV was calculated and used in all numerical data analyses. The difference was measured as the number of orchidometer sizes away from the true volume, rather than actual millilitres, since the Prader orchidometer constrains the user to select one of 12 volumes (1, 2, 3, 4, 5, 6, 8, 10, 12, 15, 20 and 25 mL). We categorised estimates into: correct; ±1; ±2; ±3 and ±>3 orchidometer sizes away from the true TV. We used responses recorded at the first assessment to assess accuracy and reliability of trained and untrained estimates. We calculated the percentage of estimates that were correct, overestimated and underestimated for both groups at each volume. We analysed differences between groups using Chi-squared and Fisher's exact tests. The mean and standard deviation (SD) of estimations in the two groups were calculated and compared by an independent t-test. Univariate analysis of covariance (ANCOVA) was used to determine if sex, age or medical school year had an effect on number of orchidometer sizes away from the true volume. The effect of repeated examination was analysed using data collected from the participants who attended for a second assessment, with the accuracy of the two groups compared by paired t-test. We used Fleiss' kappa analysis of estimates in the trained and untrained groups to measure inter-observer reliability. We calculated the number of estimates that were the same in the first and second assessment and mean (SD) of kappa scores for each participant, as measures of intra-observer reliability. We set a p-value of 0.05 to indicate significance.

Results

Participant group

Ninety students (55F) participated in the study, 46 were categorised as "trained" having participated in the workshop, and 44 categorised as "untrained". No one had previously received TV training and there were no differences in sex, age or medical school year group between the two groups. The repeat assessment was attended by 80 (89%) participants, 40 from the trained group and 40 from the untrained group.

Accuracy

Each participant estimated the volume of six testes, giving a data set of 540 estimates, 276 trained and 264 untrained. In the first assessment estimates were correct in 29% (n=81)

of the trained group compared to 27% (n=72) of the untrained group, the difference between groups was not significant (OR 1.11, 95% CI 0.76-1.61, p=0.593). Overestimates accounted for 34% in the trained and 27% in the untrained groups and underestimates for 37 and 46%, respectively. Accuracy was highest for the 20 mL testis and lowest for the 5 mL. Participants were more likely to underestimate the size of the 4 mL testis when paired with the 3 mL testis (13% of estimates were overestimates and 64% were underestimates). When paired with the 5 mL testis, participants were more likely to overestimate the 4 mL testis (41% of estimates were overestimates and 28% were underestimates) (Figure 2).

Estimates were categorised correct, one, two, three and greater than three orchidometer sizes away from the true TV for the trained and untrained groups (Figure 2, 3). Within the trained group 47% (n=130) of estimates were one orchidometer size away from true TV, significantly higher than the 34% (n=90) in the untrained group (OR 1.72; 95% CI 1.22–2.44; p=0.002). Conversely in the untrained group 11% (n=28) of estimates were three orchidometer sizes away from true TV compared to 5% (n=10) in the trained group (OR 3.15; 95% CI 1.50–6.62; p=0.002), with 5% (n=14) and 0% (n=1) greater than three sizes away, respectively (OR 15.38; 95% CI 2.01–125.00; p<0.001). The mean (±SD) difference between estimations and true TV was 0.99 (±0.82) orchidometer sizes in the trained group (n=276) and 1.35 (±1.21) in the untrained group (n=264), a mean difference of 0.36 (95% CI 0.19–0.54, p<0.001). We found no difference accuracy of estimates by sex, age or medical school year group.

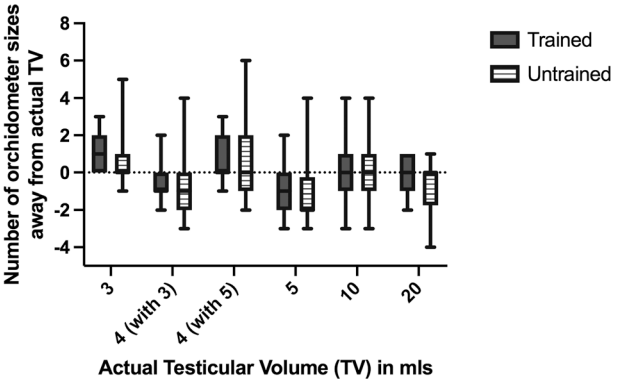

Figure 2: Accuracy of estimates in the first assessment at each volume of testis. Box and whisker plot displaying the number of orchidometer sizes away from the actual testicular volume of study participants' testicular volume estimations in the trained and untrained groups. Boxes and their central lines represent interquartile range and median values, respectively, and the whiskers indicate maximum and minimum values.

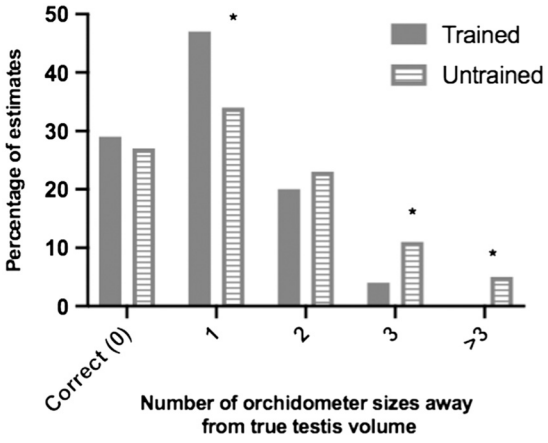

Figure 3: Overall accuracy of estimates in the first assessment (Percentage of all estimates in the first assessment±1, ±2, ±3 and ±>3 orchidometer sizes away from true TV for the trained and untrained groups. *Indicates statistically significant difference between trained and untrained groups.)

Repeat assessment

Accuracy was compared between the two assessments and between the two groups using percentage of correct estimates (Table 1). The trained group identified more TVs correctly at the second assessment (33% estimates correct vs. 28%, p<0.001) in contrast to the untrained group (25 vs. 28%, p=0.154). We found no significant difference in the mean accuracy (the number of orchidometer sizes between estimates or true TV) between the assessments in the trained and untrained groups (p=0.393 and p=0.248, respectively).

Reliability

Inter-observer reliability was assessed using Fleiss' kappa analysis. Fleiss' kappa was marginally higher for the trained group (0.16 (95% CI 0.159–0.159)) compared to the untrained group (0.11 (95% CI 0.107–0.108)). Intra-observer reliability was assessed by comparing the estimates made by volunteers in the first and second

Table 1: Intra-observer reliability. (Percentage (number) of estimates correct in watch group at the first and second assessment.).

	Accuracy in 1st assessment	Accuracy in 2nd assessment	p-Value
Trained n=240	28% (67)	33% (79)	<0.001
Untrained n=240	28% (68)	25% (59)	0.154

assessment and kappa analysis. The trained group displayed more consistency than the untrained group, with 42% (n=100) of estimates being the same as their previous estimate (irrespective of whether they were correct or not in their first estimation) in the trained group compared to 29% (n=70) in the untrained group (OR 1.74; 95%CI 1.19 and 2.53; p=0.004). Mean (SD) kappa scores of participants in the trained and untrained groups were 0.31 (0.20) and 0.19 (0.16), respectively, indicating greater concordance between estimates in the trained group compared to the untrained group.

Discussion

Overall estimation of TV by medical students was poor, with only 29 and 27% of estimates correct in the respective trained and untrained groups. Attending our workshop did not significantly impact on the number of correct estimates from participants. However, the trained group: estimated closer to the correct volume; exhibited slightly greater inter-observer and intra-observer reliability in their estimations and significantly improved their accuracy at the second attempt, where the untrained group did not.

Our results are in keeping with the findings of our previous study, using simulation to assess the accuracy and reliability of TV estimation in a large cohort (n=215) of Paediatric Endocrinologists attending an annual meeting in the UK. We reported 33% of estimations overall to be correct, 30% in those with no clinical experience and 38% in those with over 10 years of clinical experience [5]. Also consistent with our previous study, participant estimates were influenced by the size of the paired testis, conferring biological symmetry by overestimating the 4 mL testis when paired with the 5 ml testis, and underestimating it when paired with the 3 mL testis [5]. The poor intra-observer reliability resonated with our previous study with individuals agreeing with their previous estimation on only 39% of occasions [5]. Our study found underestimations of TV more common overall, contrary to other studies that have found clinicians tend to overestimate TV using the Prader orchidometer [1, 6].

The main strength of this study is the use of specifically engineered simulation models, allowing the study of significant numbers of volunteers, without the need to involve patients. Learning a practical skill requires repetition in as realistic an environment as possible. This may result in a paucity of opportunities for skill acquisition and refinement in intimate examinations, such as TV estimation, especially in children [8]. Practical examinations require a hands-on approach in order to link the theoretical knowledge learnt and the application of the required skill. Simulation offers the opportunity of unlimited repetition with immediate feedback when learning a clinical skill [14, 15]. The repeated and observed learning opportunities conducted in a safe environment also avoid the uncomfortable process of early skill acquisition on patients using trial and error [8]. The use of medical students allowed us to surpass the large recruitment target and enable the results to be extrapolated to a new trainee in paediatric endocrinology. A number of different educational modalities were employed in the workshop and assessment process, in an attempt to target different learner styles and embed knowledge and skill acquisition [12]. Although the overall accuracy was poor, training, using simulation combined with theoretical contextualisation from the video prior to the assessment, improved accuracy and reliability with more of the trained group retaining the skill at the second assessment a fortnight later.

Overall accuracy and reliability of TV estimation with the Prader orchidometer are low [5]. However, Prader orchidometers are readily available and a convenient method of assessment unlike alternatives such as ultrasound [6]. Attempts to improve TV estimations using orchidometers would therefore be of value to the paediatric endocrinology community and beyond. The results of our study indicate that accuracy and reliability, with skill retention, can all be improved using workshop style simulation training and that repetition alone, the current training method, may not lead to improvement [5]. The duration of our workshop was less than 30 min and paediatric endocrinology naïve volunteers only attended once, with no real-life experience to further embed the knowledge. Educational theory suggests that an experience needs to be repeated continuously for a skill to be learnt and the amount of repetition required is varied among individuals [13, 16]. Our previous study found no difference in accuracy of TV estimation in those with over 5 years of clinical experience, but significant improvements in those practising in the specialty for over 10 years [5]. This suggests that repetition improves accuracy but that the process takes a long time, and therefore it would be most beneficial to start any educational intervention at the beginning of an individual's training. Further research work should investigate the potential benefits of repeated workshops and assessments and skill retention over longer periods of time. This work could be extrapolated to include other pubertal staging examinations.

Simulation is not real life. The lack of penis and seminiferous tubules created an unrealistic impression

meaning our results may not be reproducible in clinical practice. The testes were slightly harder and heavier than normal for the scrotum to hang realistically. Participants had the advantage of the testes, used both in the workshop and assessment, being manufactured to match specific orchidometer sizes. In practice, testes will vary between these defined sizes. It was stressed the anatomical left and right testis should be recorded in the assessment; however, it was noted that some participants instead recorded answers according to their left and right, potentially confounding results. Fifteen participants were permitted per workshop to allow for maximum recruitment. We became aware that smaller group sizes may have been more effective, increasing the quality of the workshop through more opportunity for discussion and reflection, factors key to learning [13].

Overall accuracy and reliability of TV estimations with a Prader orchidometer were low among both trained and untrained groups. However, the trained group was more accurate and reliable and appeared to retain their newly taught skill better. Simulation should therefore be considered when teaching the difficult clinical skill of TV estimation.

Research funding: None declared.

Author contributions: CJE and NPW had the original idea for the study. BW and EW designed the workshop and undertook the pilot study. JL and DW designed and produced simulation models. NPW provided educational material for the workshops. JC, MS and CF conducted the study, analysed the data and drafted the initial manuscript. CJE revised the manuscript. All the authors have accepted responsibility for the entire content of this submitted manuscript and approved submission.

Competing interests: The funding organisation(s) played no role in the study design; in the collection, analysis and interpretation of data; in the writing of the report; or in the decision to submit the report for publication.

Statement of Ethics: Ethical approval was not required for this study. All participants recruited were willing and consented prior to taking part.

References

1. Mbaeri TU, Orakwe JC, Nwofor AM, Oranusi KC, Mbonu OO. Accuracy of Prader orchidometer in measuring testicular volume. Niger J Clin Pract 2013;16:348–51.
2. Karaman M, Kaya C, Caskurlu T, Guney S, Ergenekon E. Measurement of pediatric testicular volume with Prader orchidometer: comparison of different hands. Pediatr Surg Int 2005;21:517–20.
3. Chamberlain EN. Symptoms and signs in clinical medicine: an introduction to medical diagnosis. Baltimore: Williams and Wilkins; 1961.
4. Brunner H, Otten B. Precocious puberty in boys. N Engl J Med 1999;341:1763–5.
5. Elder CJ, Langley J, Stanton A, De Silva S, Akbarian-Tefaghi L, Wales JKH, et al. A simulation study assessing the accuracy and reliability of orchidometer estimation of testicular volume. Clin Endocrinol 2019;90:623–9.
6. Lofty-John CA, Oludayo AS, Christianah MA, Mohammed KS, Aminu MM, Mayomi O, et al. Testicular volume: correlation of ultrasonography, orchidometer and caliper measurements in children. Afr J Urol 2020;26:1–6.
7. Good M. Patient simulation for training basic and advanced clinical skills. Med Educ 2003;37:14–21.
8. Khan K, Pattison T, Sherwood M. Simulation in medical education. Med Teach 2011;33:1–3.
9. Elder CJ, De Silva S, Akbarian-Tefaghi L, Langley J, Wright NP. Inter and intra-rater reliability of accuracy of testicular volume evaluation: a simulation study. Endocr Abstr 2015;39:EP73.
10. Harold P, Mark M, Doug R, Robert B. Learning styles: concepts and evidence. Psychol Sci Public Interest 2008;9:105–19.
11. Rohrer D, Pashler H. Learning styles: where's the evidence? Med Educ 2012;46:634–5.
12. Lujan HL, Dicarlo SE. First-year medical students prefer multiple learning styles. Adv Physiol Educ 2006;30:13–6.
13. Kolb DA, Boyatzis RE, Mainemelis C. Experiential learning theory: previous research and new directions. In: Sternberg RJ, Zhang LF, editors. Perspectives on Thinking, Learning, and Cognitive Styles. New York: Routledge; 2001:227–48 pp.
14. Garner MS, Gusberg RJ, Kim AW. The positive effect of immediate feedback on medical student education during the surgical clerkship. J Surg Educ 2014. https://doi.org/10.1016/j.jsurg.2013.10.009.
15. Abrahamson SS, Denson MJ, Wolf MR. Effectiveness of a simulator in training anesthesiology residents. Acad Med 1969;44:515–9.
16. Anders Ericsson K. Deliberate practice and acquisition of expert performance: a general overview. Acad Emerg Med 2008;15:988–94.