A simulation study assessing the accuracy and reliability of orchidometer estimation of testicular volume

ELDER, Charlotte J., LANGLEY, Joe <http://orcid.org/0000-0002-9770-8720>, STANTON, Andrew, DE SILVA, Shamani, AKBARIAN-TEFAGHI, Ladan, WALES, Jerry K. H. and WRIGHT, Neil P.

Published version

ELDER, Charlotte J., LANGLEY, Joe, STANTON, Andrew, DE SILVA, Shamani, AKBARIAN-TEFAGHI, Ladan, WALES, Jerry K. H. and WRIGHT, Neil P. (2019). A simulation study assessing the accuracy and reliability of orchidometer estimation of testicular volume. Clinical Endocrinology, 90 (4), 623-629.

DR CHARLOTTE JANE ELDER (Orcid ID : 0000-0003-2390-5593)

Article type : Original Article - UK, Europe

A Simulation Study Assessing the Accuracy and Reliability of Orchidometer Estimation of Testicular Volume

Short running title: Inaccuracy of testicular volume estimation

*Charlotte J Elder [1,2], *Joe Langley [3], Andrew Stanton [3], Shamani De Silva [2], Ladan Akbarian-Tefaghi [1], Jerry KH Wales [4], Neil P Wright [2]

*authors contributed equally

[1]Department of Oncology and Metabolism, University of Sheffield, Sheffield, United Kingdom; [2]Department of Endocrinology, Sheffield Children's NHS Foundation Trust, Sheffield, United Kingdom; [3]Lab4Living: Art and Design Research Centre, Sheffield Hallam University, Sheffield, United Kingdom; [4]Univeristy of Queensland Clinical Unit, Lady Cilento Children's Hospital, South Brisbane, Queensland, Australia.

Corresponding author and contact details:

Dr Charlotte Elder

Academic Unit of Child Health,

Damer Street Building,

Sheffield Children's Hospital,

Western Bank, Sheffield

doi: 10.1111/cen.13923

S10 2TH

Email: c.j.elder@sheffield.ac.uk

Tel: 0114 305 3331

Acknowledgments: We are grateful to Nagor Ltd for their assistance in manufacturing the prosthetic testicles, Debenhams for their generous loan of the shop mannequins, Bioscientifica and the BSPED organising committee for facilitating conducting the study at the meeting, Univeristy of Sheffield medical students for their help running the study and the BSPED members who participated.

Conflicts of Interest: All authors declare there is no conflicts of interest that could be perceived as prejudicing the impartiality of this work. No specific funding was sought for this work.

Keywords: Orchidometer, measurement error, training, experience, inter-observer, intra-observer

Summary:

Context: Measuring testicular volume (TV) by orchidometer is the standard method of male pubertal staging. A paucity of evidence exists as to its inter- and intra-observer reliability and the impact of clinicians' gender, training and experience on accuracy.

Objective: Prosthetic testicular models were engineered to investigate accuracy and reliability of TV estimation.

Design: Simulation study

Setting: Conducted over three-day 2015 British Society for Paediatric Endocrinology and Diabetes (BSPED) meeting.

Participants: 215 meeting delegates (161F, 54M): 50% consultants, 30% trainees, 9% clinical nurse specialists, 11% other professionals.

Intervention: Three child-sized mannequins displayed latex scrotum containing prosthetic testicles of 3ml, 4ml, 5ml, 10ml and 20ml. Demographic data, paediatric endocrinology experience, TV examination training, examination technique and TV estimations were collected. Delegates were asked to repeat their measurements later during the meeting. Scrotum order was changed daily.

Main outcome measure: Accuracy by variance from the simulated TV. Inter- and intra-observer variability.

Results: 1284 individual estimations were obtained. Eighty-five participants repeated measurements. Delegates measured TV accurately on 33.4% (±2.6) of occasions: overestimations

37.7% (±2.3), underestimations 28.7% (±1.8) (Fleiss Kappa score 0.04). The accuracy of assessing a 4 ml testis was 36-39%. Observers underestimated the volume when paired with a 3ml testes and overestimated when paired with a 5 ml testis demonstrating a tendency impose biological symmetry. Intra-observer reliability was lacking; individuals giving different estimations for the same size testicle on 61% (±4.2) of occasions, 20% (±3.5) of estimations were more than 1 size outside the previous measurement. On only 39% (±4.2) of occasions did individuals agree with their previous estimation (irrespective of whether or not it was initially accurate). Training did not impact on results but experience did improve accuracy.

Conclusions: Overall TV estimation accuracy was poor. Considerable variation exists between and within subjects. Seniority slightly improved measurement estimation.

Introduction

Testicular volume (TV) estimation is central to the assessment of male pubertal status. Methods vary from scrotal ultrasound and calipers, to approximations using different ellipsoid calculations derived from measurements of the three dimensions of the testicle [1-5]. The most commonly used method in the paediatric outpatient setting is the Prader orchidometer, which utilises comparison of the patient's testicle to ellipsoid beads of increasing size (usually 1ml-25ml), although larger sizes may be used in andrology. Volumes 1-3ml are prepubertal, a change to a TV of 4mls indicates the onset of puberty, and adult sizes range from 12 to 25+ mls. Penile enlargement, pubic hair growth and height gain complete the constellation of puberty. Accurate confirmation of precocious puberty (TV of >3ml before the age of 9 years) is important, prompting urgent brain imaging, as a neurological pathology is the most common aetiology. The majority of testicular mass comprises of seminiferous tubules, responsible for spermatogenesis, and thus assessment of adult TV is used as a surrogate marker of function in urology and andrology.

Despite its widespread use and popularity over the last 50 years there is a paucity of research examining the accuracy and reliability of orchidometer estimations of TV. A handful of studies have been conducted, the majority assessing urologists or andrologists [6-9], but only one evaluating paediatricians [10] and three studies including prepubertal and pubertal range children [7,9,10]. The numbers of observers have been ten or fewer and inter-observer variation has been found to be high [6,8-10]. The only study reporting low inter-observer variation compared the findings of three examiners [7]. Two studies have shown good intra-observer reliability, but neither controlled for recall bias, both using examination of patients on two consecutive days [6,8]. Accuracy has been evaluated either by ultrasound, with its own incumbent variability, or displacement measurements (following bilateral orchidectomy for advanced prostatic cancer) [1-5,7-10].

We set out to evaluate the accuracy and inter- and intra-observer variability of TV estimation in a large cohort of UK-based paediatricians, using specifically engineered simulation models. Simulation enables accuracy to be established through sophisticated engineering and manufacturing

techniques, removes biological variability and allows control of factors such as ambient temperature and scrotal skin thickness. In addition, it facilitates assessment of large numbers, enabling the impact of experience, training and gender to be evaluated and a reduction of recall bias when assessing intra-observer variability.

Material and Methods

Three methodological approaches were taken for the three different stages of this research: simulation model design and development, TV assessment and data analyses.

Simulation Model Design and Development

The primary criteria were a realistic feeling to the testicular model and volume accuracy. Testicular implants are only manufactured in adult sizes (Nagor Ltd, Glasgow, UK) but offered a benchmark for synthetic texture and feel that was accepted as 'real'. The realism of the simulation models was compared to the adult implants and assessed employing an off-the-shelf industry standard, Sauter HBA 0-100 compact handheld durometer with drag indicator (from RS Components Ltd.), with a manufacturer's stated accuracy of within ±3%. This was measuring against the Shore Durometer Hardness scale (the industry standard for polymers, elastomers and rubbers)[11,12] using the measured hardness of the Nagor Testicular implant as a target value (Shore Durometer = A-05). Later this was sensed checked by qualitative feedback from 20 participants (consultant paediatric endocrinologists, endocrinology trainees and specialist nurses) in a pilot study [13]. Volume accuracy was assessed via water displacement.

The volume and radial dimension measurements were taken from Prader orchidometers and combined with the implant data. These data were used to define a parametric three-dimensional (3D) computer aided design (CAD) testis and scrotum model, which, on entering a desired TV generates the appropriately sized 3D CAD testis and scrotum. These CAD models were used as input data for Computer Numerical Control (CNC) machines to generate 3D printed casting tools and subsequent physical models. Various methods and materials were studied for casting testes to most closely approximate the texture and feel of the testicular implants using measures of Shore hardness. Following design engineering challenges such as shrinkage during curing and excessively hard testicles, the implant manufacturer, Nagor Ltd, Glasgow, UK, kindly agreed to cast the implants in silicon, to our size specifications. The volumes of the final models were all 100% accurate and the Shore Hardness values all consistent with the adult testicular implants. Qualitative feedback from clinicians confirmed that this version of the simulation models was sufficiently realistic.

3D printer moulds were used to create latex scrotum by painting the moulds with multiple layers of liquid latex, allowing each layer to cure. The scrotum were fixed to latex 'briefs' that were created

following the same process, painting them directly onto child shop mannequins. The implants were placed in the latex scrotum, with an aqueous based lubricant, sealed, and the underwear applied to the mannequins (figure 1).

The testicular volumes chosen were selected to represent both a range in size and allow closer scrutiny of the peripubertal TVs, where clinical decision making may be more critical. The 4ml testis was displayed both with a 3ml and a 5ml testis to examine how asymmetry and smaller/larger testes influence estimations. Across three child mannequins the testicular sizes were:

1. Right 3ml, left 4ml
2. Right 5ml, left 4ml
3. Right 20ml, left 10ml

Assessment

The study was conducted in November 2015, over the three-day British Society for Paediatric Endocrinology and Diabetes (BSPED) annual meeting held in Sheffield, UK. Three child-sized mannequins displayed latex scrotum containing prosthetic testes of 3ml, 4ml, 5ml, 10ml and 20ml (figure 1). The same make of orchidometer was used for all assessments (Pfizer, New York City, US). Demographic data, paediatric endocrinology experience, TV examination training and information on preferential examination technique were collected anonymously and TV estimations on all six testes were performed. Delegates were asked to repeat their measurements later during the meeting. Scrotum order was changed daily to minimise recall bias.

Data analysis

The data were entered, stored and analysed using Microsoft Excel™ and SPSS version 22. For clarity the data are presented as percentages of individuals estimating accurately, underestimating or overestimating, considered as categorical variables. Ninety–five percentage confidence intervals for the margin of error were calculated. Inter-observer reliability score was calculated using the Fleiss Kappa score as the most appropriate measure of agreement for multiple observers with categorical variables (slight agreement 0.01-0.20, fair agreement 0.21-0.40, moderate agreement 0.41-0.60, substantial agreement 0.61-0.80, almost perfect agreement 0.81-1.00)[14]. Data were also analysed by comparing TV estimations by gender, experience, occupation and training. Where participants entered a range for the testicular volume e.g. 4-5ml, rather than a discrete value, the volume closest to the actual volume was used (amounted to 2% of estimations). Data were analysed by Chi-square comparing accurate and inaccurate TV estimations by gender, experience, training and professional grade.

Results

Data were collected from 215 different delegates, 75%, were female. Consultants accounted for 50% of participants, 30% were specialist paediatric endocrinology trainees (with a minimum of seven years' post-graduate medical experience but variable paediatric endocrinology exposure), 9% endocrinology clinical nurse specialists and 11% were a mixture of recently-qualified doctors, medical students, clinical scientists and other professionals. The majority of participants (60%) had less than 5 years of experience in paediatric endocrinology, 25% had less than six months or no previous experience and 19% had worked in the specialty for more than ten years. Only 25% had received any formal training/teaching in assessing testicular volumes.

There was variation in the preferred method of assessment, with 54% examining patients recumbent and 46% standing. Approximately 65% of participants assessed TV by feeling and looking, 13% by feel alone, 19% mainly by feel and 3% mainly by looking.

The 215 participants generated 1284 individual TV estimations. Overall the correct TV was identified on 33% (margin of error 2.6%) of occasions. Overestimation by one size occurred in 25% (margin of error 2.4%) of estimations and underestimations by one size in 18% (margin of error 2.1%) (figure 2). In over a fifth of cases (22% (margin of error 1.6%)) TV was underestimated or overestimated by more than one orchidometer bead size. Overall 28% (margin of error 2.5%) of total TV estimations were underestimated and 37% (margin of error 2.6%) overestimated. Overall inter-observer reliability was poor with a Fleiss Kappa score of 0.04. A score indicating only slight agreement. There was a difference in accuracy between the various testicular volumes (table 1, figure 3). The 10 ml testis was overestimated by almost two thirds of participants and was the testis volume most likely to be inaccurately judged (83%). The accuracy of the two 4 ml testes assessments were 36-39% and there was an appreciable difference in the estimation depending on whether it was paired with a 3ml testis (40% underestimations and 21% overestimations) or a 5ml testis (13% underestimations and 51% overestimations). Individuals' ratings appear biased by a tendency to favour concordant testicular sizes and impose biological symmetry.

There were 85 individuals (510 observations) who completed the assessment on two separate occasions, allowing an estimate of intra-observer reliability. Overall on 39% (margin of error 4.2%) of occasions individuals agreed with their own previous estimation, although this estimation had not always been accurate initially. Overestimations or underestimations by one size occurred in 41% (margin of error 3.7%) of occasions and on 20% (margin of error 3.5%) of occasions individuals were more than one size outside their own previous estimation. Intra-observer agreement was poor with a Fleiss Kappa score of 0.023.

Training did not appear to make any difference to accurate TV estimations. The same percentage of participants (32%) correctly identified the volume of the testicular prosthesis with and without training. Experience did appear to improve accuracy slightly. Experience did appear to improve accuracy, but only when a threshold of years had been reached. Professionals with more than 10 years of experience in paediatric endocrinology were significantly more likely to be accurate than those with fewer than 10 years of experience (p=0.02), however there was no significant difference between those with more than 5 years of experience or those with fewer than 5 years of experience (p=0.34). Those with more than 10 years of experience correctly identified the TV 38% of times, whereas those with less than 6 months or no experience only produced correct estimations 26% and 30% of times respectively. Analysed by professional grade consultants were no more accurate in their estimations than non-consultants (p=0.13). Almost 35% of consultant estimations were correct, compared to 29% of medical students and other professionals, 32% for specialist paediatric endocrinology trainees and 27% of endocrine nurse specialists (figure 4). Gender made no difference to estimations of testicular volume.

Discussion

Overall accuracy was poor, with poor inter-observer reliability scores. The volume of the testicular prostheses correctly identified on only 33% of occasions. On 43% of occasions the TV was either underestimated or overestimated by one size, resulting in a significant discrepancy between the estimates and the actual volume in almost a quarter of TV estimations. There was a greater tendency to overestimation (37% compared to 28%), which is in keeping with previous studies [3,6-8,10]. Intra-observer variability was also significant with individuals correctly ascribing the same TV as previously on only 39% of occasions and differences of more than one size on the second estimation in 20%. The Fleiss Kappa score measuring agreement between observers and observations showed only slight agreement. Both inter- and intra-observer variability has been shown to be high in the previous studies but numbers of testes examined has been small, numbers of observers between 2 and 10, and none compared to an exact, known volume (ultrasound or post-orchidectomy water displacement were used as comparators) [6,8,10]. The only report of high levels of accuracy and correlation between and within observers studied three urologists and patients examined on two consecutive days, introducing the possibility of recall bias [7].

The greatest variance in the estimation of TV occurred with the 10 ml testicular prosthesis. There was no correlation between increasing size and improved or reduced accuracy, in contrast to studies where estimates have been reported to be more accurate at larger volumes [6,7] and *vice versa* [8]. The clinical repercussions of inaccurate evaluation alter with different sizes. A one size error at 12 or 15 mls is unlikely to have any clinical sequalae, whereas a one size error at 3 or 4 mls may have very significant implications. The definition of the onset of male puberty is a TV of 4 mls, it is therefore an important volume for clinicians to accurately assess. In this study 55% of estimations of the 3ml testis may have led to a misdiagnosis of precocious puberty, potentially yielding urgent investigations, including brain imaging, commencement of puberty-blocking medication with the associated anxiety for patient and care givers. The 4ml testis was included twice, to further examine

intra-observer reliability. Clinicians appeared influenced by whether the 4 ml testis was paired with a 5 ml or 3 ml prosthesis, underestimating when paired with the smaller 3 ml testis and overestimating with the larger 5 ml testis. In reality testes, like feet, are often not the same size, but this suggests there may be a bias when estimating testicular volume, with a tendency to confer biological symmetry. This area needs further study.

A study using simulation models of paediatric testes and scrotum for the estimation of TV has inherent limitations. Whilst significant expertise and time had been invested in the simulation attaining realism, by creating as lifelike a model as possible, there are inevitably differences between our static simulation model and real patients, such as the lack of an epididymis and penis. The study demonstrated that 54% of clinicians prefer to examine patients lying down and our model presented the "patients" standing, due to the personal examination preferences of the study investigators. It is known that there are subtle differences in the shape of the beads of different makes of Prader orchidometer (although the volume should remain constant). Using a single orchidometer to calculate the dimensions required to generate the moulds for casting the testicular prostheses and during the assessments controlled for this potential limitation. Simulation did confer considerable advantages to the study e.g. a reduction in biological variability, confidence in the accuracy of the TV being estimated and large numbers of participants.

Training in testicular examination is difficult in children and adolescents as they may be, understandably, reluctant to undergo such intimate examinations multiple times or with an audience. It was however surprising to find that 75% of the delegates reported no formal training in the assessment of TV. Practical examinations do not lend themselves easily to cognitive learning and need a "hands on" approach. Training in practical procedures and intimate examinations is a considerable challenge in medical education and simulation is increasingly utilised to bridge the gap between theoretical knowledge and practical experience [15]. Although there was no difference in the accuracy of those who had or had not previously received training, over time individuals appear to acquire some modest increase in expertise in assessing TV but training, particularly for those starting their careers, may improve their early accuracy and requires further study. A more rigorous, quality-assured, training package may also improve the accuracy of more experienced clinicians, although devising a study to demonstrate this would be challenging. Similar simulation models to those used in this study could be utilised for such further work.

Overall the assessment of testicular volume, within a simulation model, was only accurate in a third of estimations and was out by more than one size in over a fifth of cases. TVs were more likely to be over-estimated. There was significant both inter- and intra-observer variability. It is important that clinicians are aware of the extent to which their own estimates may vary from visit to visit, the tendency to over-estimate and confer biological symmetry and also the extent to which estimates by different clinicians may affect the perceived progression of puberty. Inexperienced clinicians were only slightly less accurate than those with greater experience in the specialty. We hope that the results of this study will prompt clinicians to reflect on measurement error within medicine, and

paediatric endocrinology in particular, where being aware of the flaws of a clinical tool reminds us to adopt a pragmatic approach and not depend too heavily on a single parameter. An alternative measurement technique, less prone to human error, warrants consideration, especially as important clinical decisions are made at TVs only subtly different in size to those resulting in quite different clinical pathways. Self-estimation, with boys examining their own TVs, could be considered for older patients considering the accuracy of untrained, inexperienced doctors was not vastly different from established endocrinologists. Training for all, in particular those new to the specialty, may improve the accuracy and requires further study.

Author contributions

JW had the original idea for the study. CJE, JL, AS and NPW designed the study. JL and AS designed and produced the simulation models. SDS and LAK undertook the pilot work and conducted the study. SDS, LAK, CJE and NPW analysed the data. CJE drafted the initial manuscript and all authors contributed to revising the manuscript.

References

1. Bhat S, Sathyanarayanaprasad M, Giridhar A, Srinivasa Y, Paul F. Testicular Volume Measurement: Comparison of Prader's Orchidometry, Ultrasonography, and Actual Volume by Water Displacement. *Journal of Integrative Nephrology and Andrology.* 2016;3(3):92-95.
2. Lin C-C, Huang WJS, Chen K-K. Measurement of Testicular Volume in Smaller Testes: How Accurate Is the Conventional Orchidometer? *Journal of Andrology.* 2009;30(6):685-689.
3. Mbaeri TU, Orakwe JC, Nwofor AME, Oranusi CK, Mbonu OO. Ultrasound measurements of testicular volume: Comparing the three common formulas with the true testicular volume determined by water displacement. *African Journal of Urology.* 2013;19(2):69-73.
4. Sakamoto H, Saito K, Oohta M, Inoue K, Ogawa Y, Yoshida H. *Testicular Volume Measurement: Comparison of Ultrasonography, Orchidometry, and Water Displacement.* Vol 69. Urology2007.
5. Sotos JF, Tokar NJ. Testicular volumes revisited: A proposal for a simple clinical method that can closely match the volumes obtained by ultrasound and its clinical application. *International Journal of Pediatric Endocrinology.* 2012;2012(1):17-17.
6. Carlsen E, Andersen A-G, Buchreitz L, et al. Inter-observer variation in the results of the clinical andrological examination including estimation of testicular size. *International Journal of Andrology.* 2000;23(4):248-253.
7. Karaman MI, Kaya C, Caskurlu T, Guney S, Ergenekon E. Measurement of pediatric testicular volume with Prader orchidometer: comparison of different hands. *Pediatric Surgery International.* 2005;21(7):517-520.
8. Tatsunami S, Matsumiya K, Tsujimura A, et al. Inter/intra investigator variation in orchidometric measurements of testicular volume by ten investigators from five institutions. *Asian journal of andrology.* 2006;8(3):373-378.

9. A. Diamond D, J. Paltiel H, Dicanzio J, et al. *Comparative assessment of pediatric testicular volume: Orchidometer versus ultrasound.* Vol 164. J Urol2000.

10. Rivkees SA, Hall DA, Boepple PA, Crawford JD. Accuracy and reproducibility of clinical measures of testicular volume. *The Journal of Pediatrics.* 1987;110(6):914-917.

11. Qi HJ, Joyce K, Boyce MC. Durometer Hardness and the Stress-Strain Behavior of Elastomeric Materials. *Rubber Chemistry and Technology.* 2003;76(2):419-435.

12. British Standard 903. Methods of testing vulcanised rubber Part 19 (1950) and Part A7 (1957). 1950, 1957.

13. Elder C, De Silva S, Akbarian-Tefaghi L, Langley J, Wright N. Inter and intra-rater reliability of accuracy of testicular volume evaluation: a simulation study. *Endocrine Abstracts.* 2015;39:EP73.

14. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics.* 1977;33(1):159-174.

15. Dilaveri CA, Szostek JH, Wang AT, Cook DA. Simulation training for breast and pelvic physical examination: a systematic review and meta-analysis. *BJOG.* 2013;120(10):1171-1182.

Table 1: The percentage of all participants who underestimated testicular volume by more than one Prader orchidometer size, underestimated by one size, assessed accurately, overestimated by one size and overestimated by more than one orchidometer size. Data is presented for each testicular prosthesis – 3ml, 4ml (paired with 3ml prosthesis), 4 ml (paired with 5ml prosthesis), 5ml, 10ml and 20 ml together with overall accuracy.

	Mannequin 1		**Mannequin 2**		**Mannequin 3**		**Overall Accuracy**
Testicular Volume	**3 mls**	**4 mls (paired with 3 ml prosthesis)**	**4 mls (paired with 5 ml prosthesis)**	**5 mls**	**10 mls**	**20 mls**	
Overall % underestimate	11%	40%	13%	37%	19%	43%	**28.7%**
Underestimated by more than 1 size	1% (3/214)	8% (17/214)	3% (6/214)	12% (26/214)	8% (17/214)	21% (45/214)	**10.3%**
Underestimated by 1 size	10% (20/214)	32% (68/214)	10% (21/214)	25% (54/214)	11% (23/214)	22% (47/214)	**18.4%**
Accurate	**34%** (73/214)	**39%** (84/214)	**36%** (77/214)	**31%** (67/214)	**17%** (37/214)	**39%** (84/214)	**33.4%**
Overestimated by 1 size	39% (84/214)	15% (32/214)	26% (57/214)	23% (50/214)	30% (64/214)	18% 38/214)	**25.2%**
Overestimated by more than 1 size	16% (35/214)	6% (13/214)	25% (53/214)	8% (17/214)	34% (73/214)	0% (0/214)	**12.5%**
Overall % overestimate	55%	21%	51%	31%	64%	18%	**37.7%**

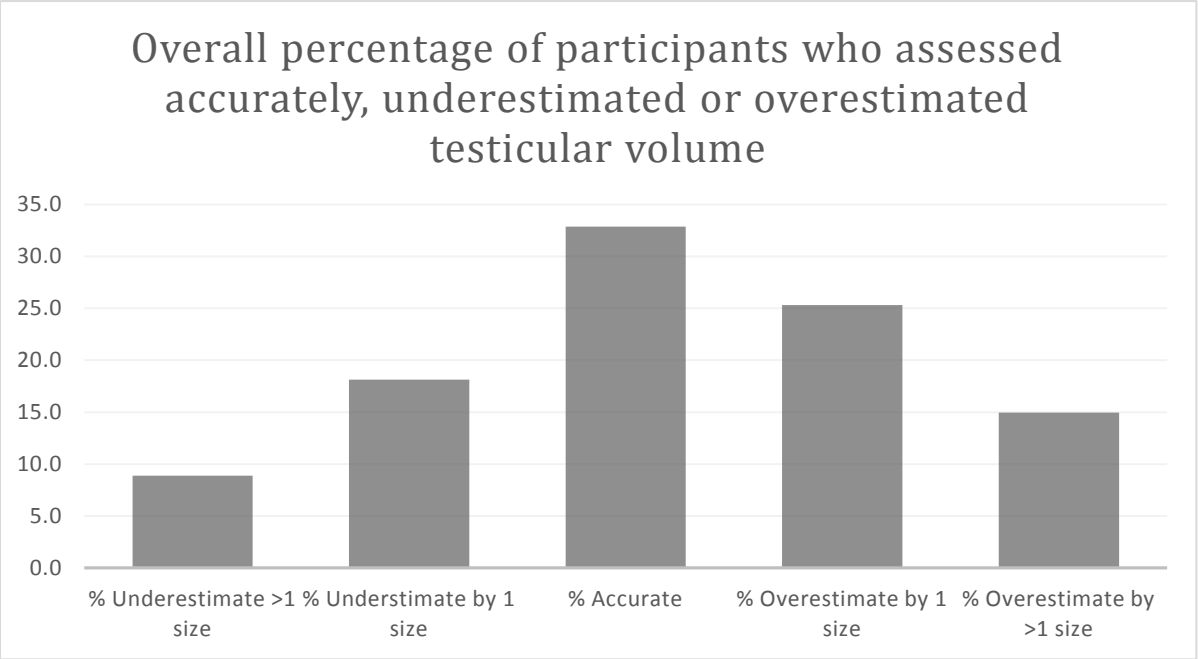
Figure 1. Exhibition stand with mannequins displaying testicular prostheses used in the simulation study. Inset shows an example of the latex "pants" and scrotum with silicon testicles.

Figure 2. Overall percentage of participants who assessed accurately, underestimated or overestimated testicular volume

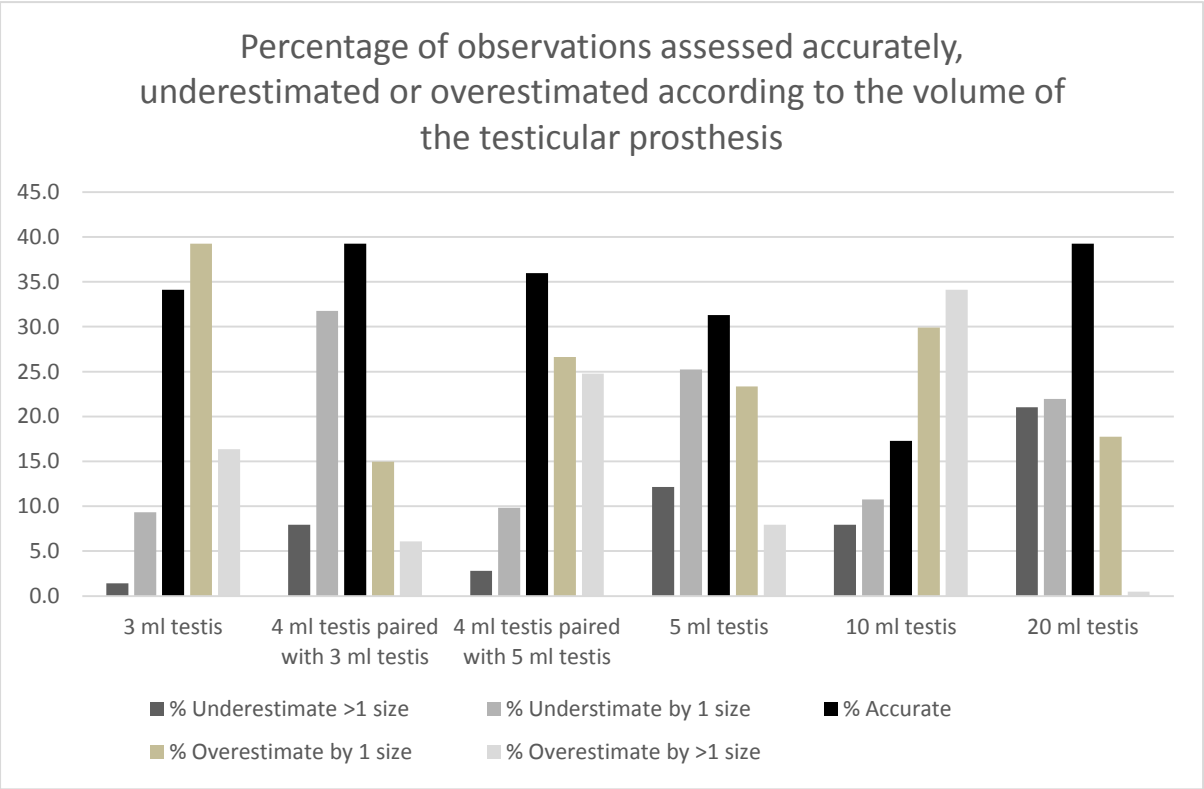
Figure 3. Percentage of observations assessed accurately, underestimated or overestimated according to the volume of the testicular prosthesis

Figure 4. Percentage of participants who overestimated, underestimated and accurately assessed testicular volumes according to Professional role

Proposed New figure 2



Proposed New figure 3



Proposed replacement Figure 2 – figure 4: