SCANDINAVIAN JOURNAL OF UROLOGY
2021, VOL. 55, NO. 5, 399–403
https://doi.org/10.1080/21681805.2021.1960599

ARTICLE

Virtual reality simulator metrics cannot be used to assess competence in ureteronephroscopy and stone removal – a validation study

Julia Dagnaes-Hansen[a,b,c], Lars Konge[a], Kim Hovgaard Andreassen[d] and Rikke Bølling Hansen[a,c,d]

[a]Copenhagen Academy for Medical Education and Simulation (CAMES), Rigshospitalet, Copenhagen, Denmark; [b]Urological Research Unit, Copenhagen University Hospital – Rigshospitalet, Copenhagen, Denmark; [c]Department of Clinical Medicine, University of Copenhagen, Copenhagen, Denmark; [d]Department of Urology, Copenhagen University Hospital – Herlev and Gentofte, Copenhagen, Denmark

ABSTRACT

Objective: The growing use of simulation-based training makes it necessary to develop efficient training programs in order to ensure optimal use of time and resources. Our aim was to develop and gather validity evidence for a simulation-based test in ureteronephroscopy and set a pass/fail standard for the test that will allow future mastery learning.

Design: This study is a validation study. A test in ureteronephroscopy and stone removal on the URO Mentor™ virtual reality simulator (3D Systems, USA) was developed by two experienced urologists in order to ensure *content*. Participants with different experience completed three standardized tasks on the simulator and simulator-generated metrics were used as outcome parameters to minimize bias and ensure a fair *response process*.

Results: Twenty novices, 15 intermediates, and 8 experienced urologists were included in the study. Validity evidence for *internal structure* and *relationship to other variables* was questionable with weak and mostly insignificant correlations across all four metrics (Cronbach's alpha = 0.14, $p = 0.15$) and across the three modules (Cronbach's alpha = 0.41 ($p = 0.02$), 0.35 ($p = 0.06$), 0.10 ($p = 0.35$), and 0.30 ($p = 0.09$) for each metric, respectively). It was not possible to establish a pass/fail score for the simulation test with meaningful *consequences*.

Conclusion: Our study showed that automatically generated simulator metrics cannot be used as a valid way of assessing competence in ureteronephroscopy. Virtual-reality simulator training could still be a valuable and patient-safe way to practice these skills, but an experienced supervisor is needed to determine when the trainee is ready to continue to supervised practice on patients.

ARTICLE HISTORY
Received 17 May 2021
Revised 12 July 2021
Accepted 20 July 2021

KEYWORDS
Ureteroscopy; virtual reality; medical education; urology urolithiasis

Introduction

In recent years, the management of upper urinary stones has shifted from shock wave lithotripsy (SWL) to endoscopic treatment [1]. This shift is also reflected in European and American guidelines on the treatment of ureteral and renal stones [2,3]. This shift demands sufficient training programs and a valid assessment of competence for urology residents.

Junior doctors in the field of urology must acquire many new skills and much time must be spent in the operating theatre. The traditional Halstedian apprenticeship model; "see one, do one, teach one" has so far been the gold standard of acquiring surgical skills. However, this type of training and apprenticeship is performed on patients and often lacks structure. Further, it can be difficult to rely on operating time for supervised training in a busy and sometimes stressful everyday setting.

Simulation-based training is the optimal way to ensure competency without endangering patient safety. A mastery learning curriculum ensures that all trainees have sufficient time to practice and learn new skills, however, it is crucial to have a valid test to ensure that all trainees have reached competency at the end of the training programme.

Traditionally, competence in surgery has been defined based on the number of procedures done. However, all trainees learn at different paces and the performance of a predefined number of performed procedures can never ensure competency [4]. In the development of a training curriculum for a specific procedure, it is important to have a final test with a credible pass/fail criterion in order to distinguish between those who have acquired the intended skills of the procedure and those who have not. Although several studies have described the effect of simulation-based training in ureteronephroscopy [1,5–9] there have been no studies describing a pass/fail to set for the acquired skills.

It is crucial to gather validity evidence for the test in order to ensure that the test measures what it is supposed to. In the last century, there has been a shift in the way of thinking validity [10–13]. At the beginning of the twentieth century, face validity, content validity, and criterion validity were the main methods to establish test validity. In 1954, construct validity was introduced and used as an overall term [14].

The modern approach of thinking validity was first presented by Messick in 1989. Messick's framework is based on five different sources of validity: content, response process,

CONTACT Julia Dagnaes-Hansen Julia.dagnaes@gmail.com Copenhagen Prostate Cancer Center, Ole Maaløes Vej 75, 7521, København N, 2200, Denmark

internal structure, relations with other variables, and consequence of the test. This framework was adopted as a standard in 1999 and confirmed again in 2014 by the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education [15].

The aim of this study was to develop and gather valid evidence for a simulation-based test of competence in ureteroscopy on the URO Mentor, (3D Systems, USA) in order to set a pass/fail criterion necessary for mastery-learning.

Materials and methods

Development of the test

The test on the URO Mentor was developed by two experienced urologists (KA and RBH) in ureteronephroscopy in order to ensure *content validity*. The test assesses skills of ureteronephroscopy and stone removal in different cases with increasing difficulty. There is accounted for both "floor" and "ceiling" effect in the development of the test by using cases with increasing difficulty.

A pilot study was carried out with one experienced urologist and one novice to clarify if the test was relevant and feasible to complete.

The test consisted of three case modules on the simulator, Stone manipulation 2 (SM2) with a distal ureteral stone, Stone manipulation 7 (SM7) with a calyceal stone in the right kidney and Stone manipulation 9 (SM9) with an obstructing mid-pole kidney stone on the left side.

In all three cases, the participants had to place a guidewire in the ureteral orifice with a rigid cystoscope, change to a ureterorenoscope and subsequently disintegrate the stone with a holmium laser. Finally, the stone should be retracted with a tip-less stone basket.

Setting and participants

The test was performed on URO Mentor from 3D Systems. The participants included novices, intermediate residents and experienced urologists in ureteronephroscopy and stone removal.

Novices were defined as medical students having no prior experience with ureteronephroscopy, intermediates were defined as residents in urology having performed between five and 50 procedures, and experienced were defined as having performed more than 100 procedures.

The participants in the group of experienced urologists and the intermediate group were from the Urological departments in Herlev and Gentofte Hospitals, Roskilde and Rigshospitalet and were invited through written invitations. The medical students in the group of novices were invited through university communication channels.

Test procedure

In order to clarify prior experience with the simulator and prior experience with endoscopy, the participants filled out a schedule of background information.

All participants completed the same warm-up session of 15 min on the URO Mentor to become familiar with the simulator. The warm-up session consisted of two modules on the simulator, Basic Task 3 and 7.

After the standardized warm-up session, the participants completed the three case modules SM 2, SM 7 and SM 9 once each. There was a time cap of 15 min for each of the three case modules.

During the test, the participants had assistance in relation to the use of guidewires, drain/flush function, pyelography, fluoroscopy and baskets to the same extent as in real-life endoscopic surgery. The assistant was one of two doctors (RBH or JD) trained in using the URO Mentor.

A sheet with a description of the three case modules (SM2, SM7 and SM9) was available to the participant throughout the entire procedure.

The assistant was not allowed to help the participant with the procedure except with the before mentioned practical help. The assistant was not allowed to give tips and tricks regarding the procedure.

In order to ensure the *response process*, all participants (novice, intermediates and experienced) were tested in the exact same way, had the same amount of warm-up time on the simulator and the same available assistance to complete the different tasks. Metrics were used in order to eliminate bias.

Data collection and analysis

All participants were given a username on the simulator with a letter categorizing the group (N for novices, I for intermediates and E for experienced) as well as a number for unique identification (i.e. N001).

All participants gave written consent that the data from the simulator was collected and used for data analysis.

Outcome measures

Based on experiences from the pilot study, the following variables were omitted from further test and analysis: x-ray time, as there is little clinical relevance of this parameter and the pilot test showed that the x-ray was only used for few seconds in each of the cases.

Total fragmentation time would be relevant in a clinical setting, but there was no difference seen in the test setting/pilot test.

Perforations were omitted from the test as there were no occurrences in neither of the groups.

In the parameter Laser work less than 3 mm from scope, the experienced urologist performed worse than the novice therefore this parameter was removed as well. The parameter Trauma from tools had a lot of missing data, and therefore this parameter was not included either.

Four variables were left for analysis: total time of procedure, trauma from scope, time to progress from orifice to pathology, and number of attempts to insert guide wire to ureteral orifice.

Statistics

IBM SPSS Statistics (SPSS) Version 25 was used for statistical analysis.

Validity evidence for the *internal structure* was explored by calculating Cronbach's alpha across all four metrics (i.e. internal consistency reliability exploring whether the four metrics measure the same trait) and for each of the metrics across all three modules (i.e. test-retest reliability exploring whether participants perform consistently from one module to the next)

The relationship to other variables (i.e. the ability of the simulator metrics to discriminate between the three groups) was examined by Analysis of Variances (ANOVA) for normally distributed data (as assessed by skew and kurtosis and Shapiro Wilk's test) and independent samples Kruskal-Wallis test for non-parametric data.

We planned to explore *Consequences* by establishing a pass/fail-standard using the contrasting groups' standard-setting method [16] and report the number of false positives (i.e. novices that pass the test) and false negatives (i.e. experienced that fail the test).

Results

A total of 44 participants participated in the study. One intermediate participant was excluded due to damage to the semirigid scope, and could therefore not complete the test, resulting in 15 intermediate participants, 20 novices, and eight experienced urologists included in this study. The demographic information on the participants can be seen in Table 1.

When looking at the four included simulator metrics across the different modules SM2, SM7 and SM9, validity evidence for the internal structure were questionable. Across all fours metrics, Cronbach's alpha was 0.14 ($p = 0.15$).

Across the three modules Cronbach's alpha for total test time was 0.41 ($p = 0.02$), for the time from orifice to pathology it was 0.35 ($p = 0.06$) and for a total number of trauma caused by the scope and a total number of attempts to place guidewire, it was 0.10 ($p = 0.35$) and 0.30 ($p = 0.09$) respectively.

Table 1. Background information on participants.

	Novices	Intermediates	Experienced
Number	20	15	8
Female gender %	70%	73%	63%
Age, mean (SD)	29.2 (1.9)	34.6 (3.9)	48 (10.9)
Years as doctor, mean (SD)	0.8 (1.2)	6.4 (3.9)	19.7 (9.4)
Number of ureteroscopies, mean (SD)			
Supervised	0	8.5 (15.3)	
Performed independently	0	12.1 (10.1)	421.3 (299.4)

Validity evidence for the *relationship to other variables* was limited (Table 2). The metrics "total time from orifice to pathology" and "total number of attempts to place guidewire" could significantly discriminate between the novices, intermediates, and experienced urologists ($p = 0.03$ and $p = 0.04$, respectively) however, the group differences in total test time and the total number of trauma caused by scope were not significant ($p = 0.10$ and $p = 0.11$, respectively).

It was not possible to establish a pass/fail score for the simulation test with meaningful *consequences*.

The scores in the different groups on the different parameters are described in Table 2.

Discussion

We here present a study where the simulator metrics from the UroMentor could not be used to assess ureteroscopy competence

We found that only a few of the metrics on the simulator could differentiate between experienced urologists and novice doctors in performing stone removal and ureterorenoscopy.

Similar results were found in a study from 2018 by Aloosh et al. [17] where 30 urology residents were tested on the URO mentor at OSCE. The participants were assessed on a Global rating scale (GRS) for ureteroscopy and the metrics from the simulator were evaluated. There was no significant difference between competent and non-competent participants when looking at the simulator metrics, however, the participants with prior training had a significantly higher GRS score than those without prior training ($p < 0.001$).

Furthermore, they found a positive correlation between previous ureteroscopy experience and trauma from the scope on the simulator. The authors discuss this as an effect of the experienced participants having more experience in the OR and therefore would be more confident on the simulator and thus less careful. This was similar to the findings of our study, where the experienced urologists had almost the same amount of trauma from the scope in SM2 as the novice group. This also supports our significant finding in the metric "total time of progressing from the orifice to pathology" in SM7, where a calyceal stone should be extracted from the right kidney. The experienced group performed significantly faster than the novices and intermediates showing confidence on the simulator based on previous experience from the OR.

Our study showed a significant difference in total time in SM2 between the three groups. However, it was the intermediate group who was faster, and not the experienced group as could be expected. This could be due to the experienced group's extensive knowledge of the procedure and wanting to perfect the procedure by for example dusting the stone.

Table 2. Simulator metrics scores for novices, intermediates, and experienced operators.

		Novices	Intermediates	Experienced	
Total test time in seconds	Mean (SD)	1957 (417)	1751 (319)	1710 (404)	$p = 0.10$
Total time from orifice to pathology in seconds	Median (range)	241 (135-534)	173 (84-335)	149 (110-254)	$p = 0.028$
Total number of trauma caused by scope	Median (range)	27 (13-230)	22 (15-30)	21 (13-29)	$p = 0.11$
Total number of attempts to place guidewire	Median (range)	8 (4-44)	5 (3-23)	5 (3-12)	$p = 0.042$

Even though our study concludes that the simulator metrics cannot be used to set a pass/fail standard, previous studies have shown an effect of training on the acquisition of stone removal skills and ureteroscopy on the URO Mentor. Wilhelm et al. [5] found that medical students who were tested before and after training on the simulator (evaluated on a GRS by two experienced urologists) had a significantly higher posttest score in the training group than those who did not train on the URO Mentor.

Virtual reality simulators can be used both for the training of skills and for the assessment of skills. It is furthermore important that simulator-based training is validated and transferred towards the actual procedure. Numerous studies have shown trainees to improve their performance in the OR after training on a simulator and a wide variety of training options and models are available, however, there are only a few transfer studies validating the transfer of simulator acquired ureteroscopy skills to the OR [18–20].

Some studies have shown simulator metrics to distinguish between novices and experienced [21] but the assessment can also be made by using OSATS or global rating scales [22,23]. This demands an experienced surgeon/physician to be present in order to determine whether the trainee has reached a sufficient level or not. The time of doctors is costly and sparse, and therefore alternative solutions can be used. A previous study has shown video-based assessment to be a reliable tool for assessing the performance in cystoscopy [24] and another study showed that medical students could assess performances as well [25].

Another argument to use the URO Mentor in a learning curriculum for urologists was presented by Dolmans et al. [26] in a study from 2009 examining the realism of the simulator. Eighty-nine urologists and residents filled out a questionnaire after having performed tasks of urethra-cystoscopy and ureterorenoscopy on the URO Mentor. They reported a high degree of realism and usefulness of the simulator. The haptic aspects of the simulator received lower ratings, which also supports our findings that the experienced urologists had equal amounts of trauma from the scope as did the novices in SM2. The lack of haptic feedback on the simulator might also explain the high amount of trauma from the scope in SM9.

One participant in the intermediate group was excluded from our study, as the participant broke the semirigid scope in SM2. This furthermore emphasizes the importance of simulator training prior to performing procedures in real life in order to ensure patient safety and longevity of the scopes.

A review by Hosny et al. regarding the durability and repair of scopes discusses that one factor in the durability of scopes is whether they are used in high volume centres or not, but the study also concludes, that guidelines on the use of ureteroscopes and simulated training can reduce the rate of ureteroscope damage [27]. Even though the technology of flexible scopes has evolved significantly, the scopes still lose deflection ability, and fibres break with the use [28]. In a study by Martin et al. looking at the economic implications of the use of reusable flexible ureteroscopes, the average time to failure of reusable scopes was 12.5 cases and resident trainees were involved in all cases where the scope was damaged [29].

Data compiled from four major ureteroscope manufacturers concluded that the most common types of damage to flexible ureteroscopes were due to working channel damage (from laser-burn or instrument passage) and scope deflection with an instrument in the working channel [30]. The same study concluded that damage to semirigid scopes was due to over-torquing which was also what we observed in our study. This further emphasizes the need for training prior to handling the scopes in an operating room, both for patient safety, but also to reduce the costs of scope repair.

Our study has several limitations. All novices were voluntary participants and it can be expected that they have an interest in surgery or urology and would therefore perform better. Another limitation was that the intermediate group has relatively low URS experience, and this could have impacted the results.

Conclusion

We found that the automatically generated simulator metrics from the UroMentor cannot be used as a valid way of assessing competence in ureteronephroscopy. Only a few of the simulator metrics were able to significantly distinguish between novices and experienced urologists. Virtual-reality simulator training could be a valuable and patient-safe way to practice these skills, but in order to set a pass/fail criterion, there must be an experienced urologist present to determine if the participant has passed or failed the test. Alternatively, video-based assessment can be used.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Julia Dagnaes-Hansen http://orcid.org/0000-0002-1258-5822

References

[1] Villa L, Şener TE, Somani BK, et al. Initial content validation results of a new simulation model for flexible ureteroscopy: the key-box. J Endourol. 2017;31:72–77.
[2] Assimos D, Krambeck A, Miller NL, et al. Surgical management of stones: American urological association/endourological society guideline, part I. J Urol. 2016;196(4):1153–1160.
[3] EAU guidelines on Urolithiasis. 2021. https://uroweb.org/wp-content/uploads/EAU-Guidelines-on-Urolithiasis-2021.pdf.
[4] Barsuk JH, Cohen ER, Feinglass J, et al. Residents' procedural experience does not ensure competence: a research synthesis. J Grad Med Educ. 2017;9:201–208.
[5] Wilhelm DM, Ogan K, Roehrborn CG, et al. Assessment of basic endoscopic performance using a virtual reality simulator. J Am Coll Surg. 2002;195(5):675–681.
[6] Jacomides L, Ogan K, Cadeddu JA, et al. Use of a virtual reality simulator for ureteroscopy training. J Urol. 2004;171(1):320–323.
[7] Knoll T, Trojan L, Haecker A, et al. Validation of computer-based training in ureterorenoscopy. BJU Int. 2005;95(9):1276–1279.

[8] Watterson JD, Beiko DT, Kuan JK, et al. Randomized prospective blinded study validating acquistion of ureteroscopy skills using computer based virtual reality endourological simulator. J Urol. 2002;168(5):1928–1932.
[9] Brehmer M, Tolley DA. Validation of a bench model for endoscopic surgery in the upper urinary tract. Eur Urol. 2002;42(2): 175–180.
[10] Colliver JA, Conlee MJ, Verhulst SJ. From test validity to construct validity … and back? Med Educ. 2012;46(4):366–371.
[11] Downing SM, Yudkowsky R. Assessment in Health Professions Education. Chapter 2. 2009. https://doi.org/10.4324/9780203880135
[12] Noureldin YA, Lee JY, McDougall EM, et al. Competency-based training and simulation: making a "valid" argument. J Endourol. 2018;32(2):84–93.
[13] Noureldin YA, Sweet RM. A call for a shift in theory and terminology for validation studies in urological education. J Urol. 2018; 199(3):617–620.
[14] Cronbach LJ, Meehl PE. Construct validity in psychological tests. Psychol Bull. 1955;52(4):281–302.
[15] AERA, APA and NCME. The standards for educational and psychological testing. Washington D.C: AERA publications. 2014.
[16] Jørgensen M, Konge L, Subhi Y. Contrasting groups' standard setting for consequences analysis in validity studies: reporting considerations. Adv Simul. 2018;3:5.
[17] Aloosh M, Couture F, Fahmy N, et al. Assessment of urology postgraduate trainees'competencies in flexible ureteroscopic stone extraction. Can Urol Assoc J. 2018;12(2):52–58.
[18] Seymour NE, Gallagher AG, Roman SA, et al. Virtual reality training improves operating room performance - results of a randomized, double-blinded study. Ann Surg. 2002;236(4):458–464.
[19] Childs BS, Manganiello MD, Korets R. Novel education and simulation tools in urologic training. Curr Urol Rep. 2019;20: 81.
[20] Aloosh M, Noureldin YA, Andonian S. Transfer of flexible ureteroscopic Stone-Extraction skill from a virtual reality simulator to the operating theatre: a pilot study. J Endourol. 2016;30(10): 1120–1125.
[21] Bajka M, Tuchschmid S, Fink D, et al. Establishing construct validity of a virtual-reality training simulator for hysteroscopy via a multimetric scoring system. Surg Endosc. 2010;24(1):79–88.
[22] Moktar J, Bradley CS, Maxwell A, et al. Skill acquisition and retention following simulation-based training in pavlik harness application. J Bone Joint Surg. 2016;98(10):866–870. Vol.
[23] Malacarne DR, Escobar CM, Lam CJ, et al. Teaching vaginal hysterectomy via simulation : Creation and validation of the objective skills assessment tool for simulated vaginal hysterectomy on a task trainer and performance among different levels of trainees. Female Pelvic Med Reconstr Surg. 2019;25: 300–302.
[24] Dagnaes-Hansen J, Mahmood O, Bube S, et al. Direct *Observation vs. Video-Based Assessment in Flexible Cystoscopy*. J. Surg. Educ. 2018;75(3):671–677.
[25] Mahmood O, Dagnaes J, Bube S, et al. Nonspecialist raters can provide reliable assessments of procedural skills. J Surg Educ. 2018;75(2):370–376.
[26] Dolmans VEMG, Schout BMA, De Beer NAM, et al. The virtual reality endourologic simulator is realistic and useful for educational purposes. J Endourol. 2009;23(7):1175–1181.
[27] Hosny K, Clark J, Srirangam SJ. Handling and protecting your flexible ureteroscope: how to maximise scope usage. Transl Androl Urol. 2019;8(Suppl 4):S426–S435.
[28] Traxer O, Dubosq F, Jamali K, et al. New-generation flexible ureterorenoscopes are more durable than previous ones. Urology. 2006;68(2):276–279.
[29] Martin CJ, McAdams SB, Abdul-Muhsin H, et al. The economic implications of a reusable flexible digital ureteroscope: a cost-benefit analysis. J Urol. 2017;197(3 Pt 1):730–735.
[30] Sung JC, Springhart WP, Marguet CG, et al. Location and etiology of flexible and semirigid ureteroscope damage. Urology. 2005; 66(5):958–963.