0022-5347/04/1725-1953/0
THE JOURNAL OF UROLOGY®

Vol. 172, 1953–1957, November 2004
Printed in U.S.A.
DOI: 10.1097/01.ju.0000141298.06350.4c

FACE, CONTENT AND CONSTRUCT VALIDITY OF THE UNIVERSITY OF WASHINGTON VIRTUAL REALITY TRANSURETHRAL PROSTATE RESECTION TRAINER

R. SWEET,* T. KOWALEWSKI, P. OPPENHEIMER,† S. WEGHORST† AND R. SATAVA

From the Departments of Urology (RS) and General Surgery (RS), University of Washington and Human Interface Technology Laboratory (RS, TK, PO, SW), Seattle, Washington

ABSTRACT

Purpose: We examined the face, content and construct validity of version 1.0 of the University of Washington transurethral prostate resection (TURP) trainer.

Materials and Methods: Version 1.0 of a virtual reality based simulator for transurethral skills was developed at our laboratory by integrating TURP hardware with our virtual 3-dimensional anatomy, irrigation control, cutting, bleeding and haptics force feedback. A total of 72 board certified urologists and 19 novices completed a pre-task questionnaire, viewed an introductory training video and performed a pre-compiled 5-minute resection task. The simulator logged operative errors, gm resected, blood loss, irrigant volume, foot pedal use and differential time spent with orientation, cutting or coagulation. Trainees and experts evaluated the simulator on a modified likert scale. The 2-tailed Levene t test was used to compare means between experts and novices.

Results: Overall version 1.0 content was between slightly and moderately acceptable. Experts spent less time with orientation ($p < 0.0001$), resected more total tissue ($p < 0.0001$), had more gm resected per cut ($p = 0.002$) and less blood loss per gm resected ($p = 0.032$), used less irrigant per gm resected ($p = 0.02$) and performed fewer errors ($p < 0.0001$) than novices.

Conclusions: We established the face, content and construct validity for version 1.0 of the University of Washington TURP trainer to simulate the skills necessary to perform TURP. A predictive validity study showing a translation of skills from the virtual environment to the operating room will complete the validation of this model.

KEY WORDS: prostate, transurethral resection of prostate, computer simulation, user-computer interface, medical education

Urology is a technically oriented field. Although urological training is historically effective, it is expensive and lacks a standardized curriculum. Currently the field imposes virtually no objective assessment of technical skills during training, and certification and accreditation indirectly depend on the number of cases performed, the accreditation status of the residency program, and the results of written and oral board examinations.

Surgical simulation technology provides a novel solution that, when properly embedded into the curriculum, may efficiently train and accurately assess the acquisition of many skills. Given the current state of technology, endoscopic, laparoscopic and Seldinger technique procedures are amenable to simulation. Some urology specific skill simulators have been built to train cystoscopy,[1] transurethral resection of the prostate (TURP),[1–4] ureteroscopy,[5,6] percutaneous access (PERC Mentor, Simbionix USA Corp., Cleveland, Ohio) and digital rectal examinations.[7] However, only a few preliminary validation studies have been published,[8–11] including those of the URO Mentor (Simbionix)[7] and the digital rectal examination simulator.

In urology TURP represented a skill set that met all of the criteria of a good model for simulation. It is difficult to learn and teach, it relies primarily on visual cues, it remains a mainstay operation for a common medical problem (bladder outlet obstructive symptoms) and it proves highly amenable to simulation. Meanwhile, the number of TURPS being done in residency may be inadequate to train the skills required to perform the procedure.[4,12] We created a virtual reality based TURP simulator to train and assess the skills necessary to perform this procedure in a logical stepwise fashion.[4] In this study we examined the face, content and construct validity of version 1.0 of this trainer to simulate the skills necessary to perform TURP.

Gallagher et al described certain recommendations by which surgical simulators should be validated.[13] 1) Face validity is a type of validity that is assessed by having experts review the contents of a test to see if it seems appropriate. Simply stated, it is an evaluation to determine whether the test measures what it is supposed to measure. It is a subjective type of validation and it is usually only used during the initial phases of test construction. 2) Content validity is an estimate of the validity of a testing instrument based on a detailed examination of the contents of the test items. Evaluation is done by reviewing each item to determine whether it is appropriate to the test and by assessing the overall cohesiveness of the test items, such as whether the test contains the steps and skills that are used in a procedure. Establishing content validity is also a largely subjective operation and it relies on the judgments of experts concerning the relevance of the materials used. 3) Construct validity is a set of procedures for evaluating a testing instrument based on the degree to which the test items identify the quality, ability or trait that it was designed to measure. As more traits or performance qualities are identified, construct va-

Accepted for publication June 4, 2004.
Study received human subjects committee approval.
Supported by the American Foundation for Urologic Disease/American Urological Association funded Research Scholar Program, Mary Gates Foundation and ACMI Corp.
* Correspondence: University of Washington Urology, Mail Stop 356510, Seattle, Washington 98195-6510 (telephone: 206-616-1573; FAX: 206-543-3272; e-mail: rsweet@u.washington.edu).
† Financial interest and/or other relationship with ACMI, Inc.

lidity must be updated. A common example is the ability of an assessment tool to differentiate experts and novices performing a given task. This is the construct that we examined in this study. Concurrent validity, discriminate validity and predictive validity were not addressed in this study.

MATERIALS AND METHODS

General specifications of the simulator.[4] Version 1.0 of the University of Washington TURP simulator was designed at our laboratory (fig. 1). It integrates our novel 3-dimensional virtual anatomy with a physical model (Simulab, Seattle, Washington). Tracking of the camera and loop is provided by devices (Polhemus, Colchester, Vermont, and Mimic Technologies, Seattle, Washington). The simulator has interactive inflow and outflow irrigation control, Mimic real-time haptic force feedback, simulated tissue cutting and bleeding controlled by a foot pedal (Bovie, St. Petersburg, Florida) with interactive sound cues. We integrated our hardware solutions and software interface with a TNGIII interface box (MindTel, Syracuse, New York) and a 27.6Fr resectoscope with an Iglesias working element (ACMI, Scarborough, Massachusetts).[4] We calibrated the simulator code for determining weight in gm by correlating the tissue weight of a single expert resection during the first 5 minutes of resection on a similar sized gland with the expert performance during the first 5 minutes on the simulator. We calibrated the irrigation rate based on inflow and outflow curves of a similar resectoscope through a single patient undergoing cystoscopy with a uroflowmetry collection device brought into the operating room. After we confirmed that content validity was established for the simulator ability to simulate bleeding, bleeding rate coefficient calibration was based on the mean expert gm per minute on the simulator and published studies reporting expert blood loss per gm estimates of 4.65 gm hemoglobin per gm resected.[14–16]

Study design. At the 2002 American Urological Association meeting 91 subjects completed the study protocol, consisting of 72 experts (board certified urologists) and 19 novices with a master level education or above. The general database was generated with a pre-task questionnaire, which provided demographic data including training status and TURP related questions. All subjects viewed an introductory training video and performed the same pre-compiled 5-minute resection task. Subjects were given a 100 gm prostate to resect and they were given 5 minutes to "resect as much tissue as possible, most efficiently, with the least amount of blood loss, using the least amount of irrigant and coagulation current." The simulator logged all features of instrument interaction with the virtual environment (fig. 1). Only data on subjects who completed the task were included in the analysis. For study purposes demographic data on training status and primary metrics consisted of operative errors, gm resected, blood loss, irrigant volume, foot pedal use, and differential time spent with orientation, cutting and coagulation. Operative errors were defined and identified as capsular perforation, external sphincter resection, rectal perforation, dorsal venous complex resection, ureteral orifice resection and undermining of the bladder neck. Three trained nonmedical technicians administered the task to the participants. They were coached beforehand by the principal investigator to give consistent, predetermined responses to anticipated questions or comments. Upon completion of the task the expert subjects completed a post-procedure questionnaire critiquing the performance, face and content validity of the simulator.

Construct validity for this tool to simulate TURP was examined using the Levene 2-tailed t test in SPSS (SPSS, Chicago, Illinois), which compared mean performance metrics for the novice and expert groups. Resident data was not included in this analysis because it was inappropriately powered to be meaningful. The mean performance metrics of experts were used as a basis to introduce the concept of criterion levels for specific skills. The study met with full approval of the American Urological Association and the University of Washington human subjects committee.

RESULTS

Demographics. All novice subjects had a masters level or higher education. Subjects were 23 to 68 years old (mean age ± SD 40.5 ± 10.6). Of these subjects 46% reported that they were in academic practice, 19% were in solo private practice, 18% were in small group practice, 10% were in large group practice and 4% worked for health maintenance organizations. With regard to place of residency 53% of the participants were trained in North America, 14% were trained in Central or South America, 26% were trained in Europe and 6% were trained in Asia. Approximately half of the participants completed training prior to 1994. The number of TUR procedures performed in the preceding month by experts was 0 to 30 with approximately half of them reporting having performed 4 or fewer procedures. The number performed in the preceding year was 0 to 500 with approximately half of them reporting 20 or fewer procedures. The number performed in the previous 2-year period was 0 to 1,000 with approximately half of them reporting 40 or fewer procedures. With regard to video game experience 47% of the participants reported never playing, 25% reported playing once yearly, 18% played monthly, 9% played weekly and approximately 1% played daily. As expected, video game experience significantly correlated negatively with participant age (Pearson $R = -0.212$, $p = 0.011$).

Face and content validity. Figure 2 shows expert data on pre-task attitudes about TURP and simulation. Figure 3 shows post-task questionnaire results. Figure 4 shows a modified 5-point Likert scale revealing that urologists ranked each of its components above the acceptability threshold. Overall participating urologists and trainees believed that version 1.0 of the UW TURP simulator was acceptable.

Criterion levels. For use in future studies the scores of accredited urologists would determine benchmark levels based on the overall means in the recorded metrics. Table 1

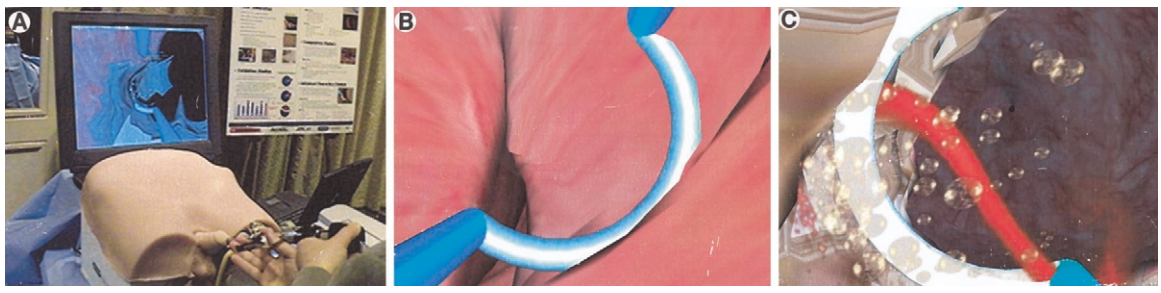

FIG. 1. *A*, University of Washington TURP simulator. *B*, endoscopic view of simulator demonstrates deformation. *C*, endoscopic view of simulator.

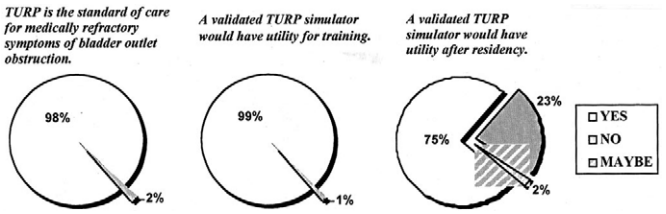

FIG. 2. Pre-task attitudes regarding simulator

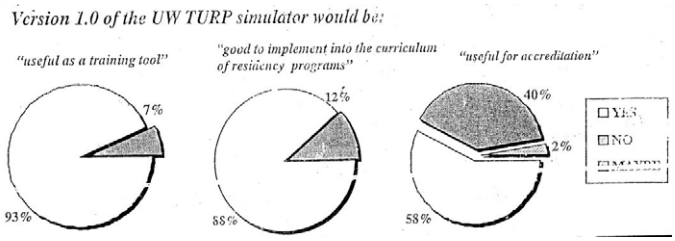

FIG. 3. Post-task attitudes regarding simulator

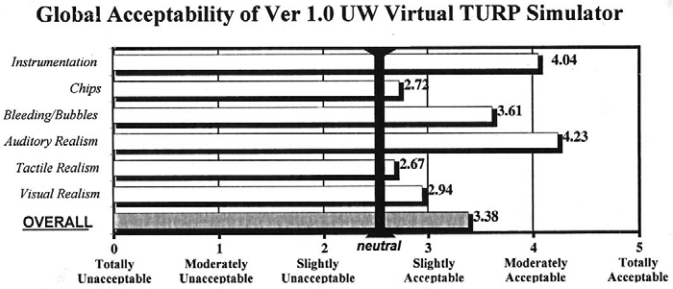

FIG. 4. Board certified urologists ranked simulator version 1 above acceptability thresholds on modified Likert scale.

lists the mean scores of board certified urologists (experts) for select metrics.

Construct validity. Experts outperformed novices in all aspects of the task. Most distinctive were the metrics and operative errors, which clearly separated the 2 subgroups (table 2, fig. 5). For the 5-minute resection task no experts or trainees logged any serious operative errors, while a significant number of errors were performed by novices (table 2).

DISCUSSION

Face validity: at face value does the TURP simulator look and perform like TURP and is there a need for such a training tool? As figure 2 shows, board certified urologists believed that TUR skills remain important and simulation should have an important role in the training of skills related to

TABLE 2. *Errors performed by 19 novices on TURP simulator*

Error	Mean	SD	SEM
Sphincter resection	0.47	0.51	0.12
Excessive bleeding	0.16	0.37	0.09
Capsular perforation	0.05	0.23	0.05

The 72 experts performed no errors.

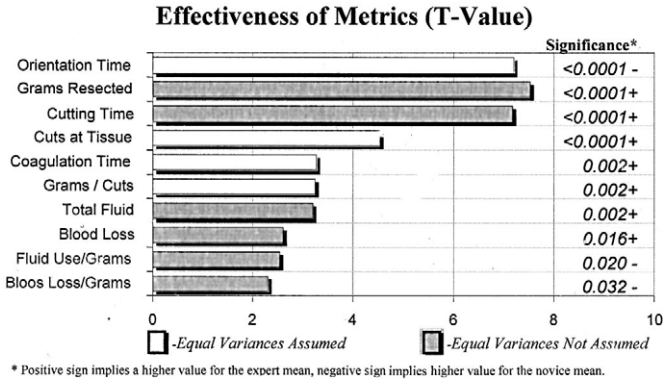

FIG. 5. Results of Levene conditioned 2-tailed t test quantifying resolving potential of select metrics to determine differences between experts and novices.

TURP. We acknowledge that there may be a selection bias toward such a response because subjects who chose to use the simulator and participate in the study were probably more likely to have such an attitude. Nonetheless, we were impressed by the overwhelming demonstrated need for a simulator to train TUR skills during residency as well as the overwhelming acceptance of the simulator to serve this purpose. Of the respondents 75% believed that a validated simulator could be used to maintain skills after residency.

Content validity: do experts believe that the specific content of the simulator is adequate to simulate and train TURP? Of the experts 93% believed that simulator version 1.0 would be useful for training TURP, while 88% thought that it was ready in its current state and should be implemented into the curriculum of residency programs. Figure 4 demonstrates what experts thought about specific aspects of the simulator. All aspects of the simulator were acceptable, although it was evident that our original force feedback device (tactile realism), visual realism and interaction with cutting/chips were areas on which to focus our efforts for version 2.0. In version 2.0, which was completed at the time of this study, we subjectively improved the interactivity and fidelity of cutting/chips, haptics force feedback and visual realism.

Construct validity. For construct validity purposes we compared novices with experts, who are presumably on the flat end of the learning curve. We did not compare trainees with novices or experts because they represent the learning curve

TABLE 1. *Group statistics for key metrics*

	19 Novices			72 Experts		
	Mean	SD	SEM	Mean	SD	SEM
Orientation time (sec)	260.85	26.37	6.05	206.60	29.86	3.52
Gm resected	2.17	1.66	0.38	6.01	2.90	0.34
Cutting time (sec)	22.05	19.25	4.42	61.04	26.86	3.17
Total cuts	51.32	32.18	7.38	81.24	23.68	2.79
Coagulation time (sec)	17.10	13.63	3.13	32.36	19.06	2.25
Gm/total cuts	0.05	0.03	0.01	0.08	0.03	0.00
Total fluid (ml)	1,833.71	814.53	186.86	2,699.95	1,664.73	196.19
Cut pedal presses	24.63	19.48	4.47	36.25	15.10	1.78
Blood loss (ml)	74.30	64.72	14.85	115.42	45.52	5.37
Fluid use (ml)/gm	1,553.72	1,771.03	406.30	515.00	467.29	52.57
Blood loss (ml)/gm	37.58	27.48	6.31	22.63	13.01	1.53
Gm/cut pedal presses	0.12	0.11	0.03	0.17	0.06	0.01
Fluid use (ml)/blood loss (ml)	66.96	102.28	24.11	26.86	18.91	2.23

and larger numbers of residents grouped by experience level would be required to do such a sophisticated analysis.

What factors distinguish novices from experts? The simulator effectively was able to distinguish novices from experts (table 2, fig. 5). As one would expect in the operating theater, the largest difference noted between novices and experts was that novices spent much more time orienting themselves within the anatomy ($p = 0.00001$), while experts spent more time cutting and coagulating ($p = 0.00001$). Experts had more total blood loss but, since blood loss highly correlated with gm resected, we also looked at blood loss per gm resected. Experts had significantly less blood loss per gm resected than novices ($p = 0.03$). Experts also used less irrigation fluid per gm resected than novices ($p = 0.02$). The ultimate goal of training is to decrease the number of errors performed in patients. None of our experts had an operative error during the 5-minute resection task but a substantial number of errors were logged among the novice group. Approximately half of the novices resected the sphincter and 16% had to stop the task secondary to red out. It did not surprise us that no experts resected the sphincter or had to stop the task because of red out but the fact that no expert even had a perforation of capsule was suspect. This can be explained by the fact that the prostate model used in the simulator was approximately 100 gm and with trained systematic resection capsular perforation was unrealistic to achieve in the 5-minute resection time allotted. Although 23% of the novices perforated the capsule, this was done near the apex, where there was less surrounding tissue. We suspect that most experts were trained to do apical dissection near the end and, because of the limitation of a 5-minute resection, experts did not get to apical dissection. In retrospect we should have used a smaller gland better to distinguish cutting skill, as defined by a minimal surface area of perforation. However, we can infer by these results that it was the expert training on real procedures that allowed them to out perform the novices.

Do the metrics measured by the simulator mimic what occurs in the operating room? This is best addressed by looking at the content validity data and expert performances (table 1, fig. 3). The trend that the amount of gm resected correlated with blood loss in the operating room is also seen as a positive correlation on the simulator. Gm resected also correlated with total cuts at the tissue and it had a negative correlation with time spent with orientation in the expert group (all $p < 0.0001$). It has been estimated that the average resection time for a skilled resectionist in the operating room is just greater than 1 gm per minute. Our experts resected an average of 6.01 gm in 5 minutes, which is consistent with this estimation. Published studies reporting expert blood loss per gm estimates and irrigant use of 4.65 gm hemoglobin per gm resected[14–16] then allowed us to calibrate coefficients for rates of simulator blood loss and fluid. It was only with the establishment of content validity that we were comfortable performing this calibration.

Criterion levels: if we were to integrate the simulator into a curriculum, what are adequate scores? When considering the introduction of a simulator into a curriculum, a consensus panel of experts must decide what the criterion levels are for a given metric necessary to pass a given skill set. Expert means or a SD could be used as an approximate target for trainees to pass on the TURP simulator after it is refined (table 1). The primary reason that we say approximate is that we are first assuming that board certification makes an expert a true expert, which may not necessarily be the case on a skill set by skill set basis. Also, we believe that there is probably a short learning curve on the simulator and the study design was such that subjects were only allowed 1 try on the simulator without any practice, which may have artificially lowered the bar on all metrics. Ideally establishing reliability and subsequent learning curves specific to the simulator first by having numerous true experts perform multiple procedures on the simulator until the metrics scores flatten out would achieve criterion levels and provide guidelines for integration into the curriculum of residency programs. Unfortunately in the pre-simulation era there are few objective ways to establish who these technical experts truly are. In a sense we are looking for a gold standard that may not exist.

Should the TURP simulator be used for accreditation? More than half (58%) of the respondents believed that the simulator would be useful for accreditation. Although this appears to be possible, it is our opinion that given the enormous implications of such an endeavor (ie exclusion of a candidate from the chosen career, potential remedial training or termination of residents from residency programs and the possibility of senior surgeons losing accreditation or stopping practice secondary to the results of this test of technical skill) "the issue of validation and the scientific integrity of these studies must be beyond reproach."[13] We believe that the first step in validating simulators for training are studies designed to establish these measures of validity, establishing reliability as well as the fact that performance on the simulator can predict performance in the operating room (predictive validity).[17] After this has been accomplished the simulator is valid for training but we would still caution its use for assessment. It is only after extensive data collection during training correlated with basic measures of performance testing and a link to improved patient outcomes that simulator usage for the purposes of baseline skills assessment should be considered.

CONCLUSIONS

We established the face, content and construct validity for version 1.0 of the University of Washington TURP trainer to simulate the skills necessary to perform TURP. A predictive validity study showing a translation of skills from the virtual environment to the operating room will complete the validation of this model. Due to the limited number of urology residents in a given program the study design would have to be of a multi-institutional nature. Based on this study we believe that integration of this simulator into the urology curriculum of residency programs for training purposes is appropriate. We caution its use or the use of any simulator for assessment and/or continuing medical education until more rigorous validation is completed.

A. Gallagher assisted with study design, M. Mayo and B. Joyner provided critical reviews, J. Berkley assisted with the force feedback device and C. Toly designed the prosthetic device.

REFERENCES

1. Manyak, M. J., Santangelo, K., Hahn, J., Kaufman, R., Carleton, T., Hua, X. C. et al: Virtual reality surgical simulation for lower urinary tract endoscopy and procedures. J Endourol, **16:** 185, 2002
2. Gomes, M. P., Barrett, A. R., Timoney, A. G. and Davies, B. L.: A computer-assisted training/monitoring system for TURP structure and design. IEEE Trans Inf Technol Biomed, **3:** 242, 1999
3. Oppenheimer, P., Gupta, A., Weghorst, S., Sweet, R. and Porter, J.: The representation of blood flow in endourologic surgical simulations. Stud Health Technol Inform, **81:** 365, 2001
4. Sweet, R., Porter, J., Oppenheimer, P., Hendrickson, D., Gupta, A. and Weghorst, S.: Simulation of bleeding in endoscopic procedures using virtual reality. J Endourol, **16:** 451, 2002
5. Michel, M. S., Knoll, T., Kohrmann, K. U. and Alken, P.: The URO Mentor: development and evaluation of a new computer-based interactive training system for virtual life-like simulation of diagnostic and therapeutic endourological procedures. BJU Int, **89:** 174, 2002
6. Preminger, G. M., Babayan, R. K., Merril, G. L., Raju, R.,

Millman, A. and Merrill, J. R.: Virtual reality surgical simulation in endoscopic urologic surgery. Stud Health Technol Inform, **29:** 157, 1996
7. Burdea, G., Patounakis, G., Popescu, V. and Weiss, R. E.: Virtual reality-based training for the diagnosis of prostate cancer. IEEE Trans Biomed Eng, **46:** 1253, 1999
8. Watterson, J. D., Beiko, D. T., Kuan, J. K. and Denstedt, J. D.: A randomized prospective blinded study validating acquisition of ureteroscopy skills using a computer based virtual reality endourological simulator. J Urol, **168:** 1928, 2002
9. Wilhelm, D. M., Ogan, K., Roehrborn, C. G., Cadeddu, J. A. and Pearle, M. S.: Assessment of basic endoscopic performance using a virtual reality simulator. J Am Coll Surg, **195:** 675, 2002
10. Jacomides, L., Ogan, K., Cadeddu, J. A. and Pearle, M. S.: Use of a virtual reality simulator for ureteroscopy training. J Urol, **171:** 320, 2004
11. Johnson, D. B., Kondraske, G. V., Wilhelm, D. M., Jacomides, L., Ogan, K., Pearle, M. S. et al: Assessment of basic human performance resources predicts the performance of virtual ureterorenoscopy. J Urol, **171:** 80, 2004
12. Holtgrewe, H. L.: Editorial: surgical management of benign prostatic hyperplasia in 2001—a pause for thought. J Urol, **166:** 177, 2001
13. Gallagher, A. G., Ritter, E. M. and Satava, R. M.: Fundamental principles of validation, and reliability: rigorous science for the assessment of surgical education and training. Surg Endosc, **17:** 1525, 2003
14. Ekengren, J. and Hahn, R. G.: Blood loss during transurethral resection of the prostate as measured by the HemoCue photometer. Scand J Urol Nephrol, **27:** 501, 1993
15. Ala-Opas, M. Y. and Gronlund, S. S.: Blood loss in long-term aspirin users undergoing transurethral prostatectomy. Scand J Urol Nephrol, **30:** 203, 1996
16. Donohue, J. F., Sharma, H., Abraham, R., Natalwala, S., Thomas, D. R. and Foster, M. C.: Transurethral prostate resection and bleeding: a randomized, placebo controlled trial of role of finasteride for decreasing operative blood loss. J Urol, **168:** 2024, 2002
17. Seymour, N. E., Gallagher, A. G., Roman, S. A., O'Brien, M. K., Bansal, V. K., Andersen, D. K. et al: Virtual reality training improves operating room performance: results of a randomized, double-blinded study. Ann Surg, **236:** 458, 2002