Development and validation of the TOCO–TURBT tool: a summative assessment tool that measures surgical competency in transurethral resection of bladder tumour

ANNA H. DE VRIES [1]; ARNO. M. M. MUIJTJENS [2;] HILDE G. J. VAN GENUGTEN [1]; AD. J. M. HENDRIKX[1]; EVERT L. KOLDEWIJN[1 3]; BARBARA M. A. SCHOUT [4 5] CEES P. M. VAN DER VLEUTEN [2] ; CORDULA WAGNER [56]; IRENE M. TJIAM [7]; JEROEN J. G. VAN MERRIËNBOER [3]

[1.]Department of UrologyCatharina HospitalEindhovenThe Netherlands
[2.]Department of Educational Development & Research, Faculty of Health, Medicine and Life SciencesMaastricht UniversityMaastrichtThe Netherlands
[3.]School of Health Professions EducationMaastricht UniversityMaastrichtThe Netherlands
[4.]Department of UrologyAlrijne HospitalLeidenThe Netherlands
[5.]Netherlands Institute for Health Services Research (NIVEL)UtrechtThe Netherlands
[6.]Department of Public and Occupational HealthEMGO Institute for Health and Care ResearchAmsterdamThe Netherlands
[7.]Department of UrologyCanisius Wilhelmina HospitalNijmegenThe Netherlands

ABSTRACT

Background

The current shift towards competency-based residency training has increased the need for objective assessment of skills. In this study, we developed and validated an assessment tool that measures technical and non-technical competency in transurethral resection of bladder tumour (TURBT).

Methods

The 'Test Objective Competency' (TOCO)–TURBT tool was designed by means of cognitive task analysis (CTA), which included expert consensus. The tool consists of 51 items, divided into 3 phases: preparatory (n = 15), procedural (n = 21), and completion (n = 15). For validation of the TOCO–TURBT tool, 2 TURBT procedures were performed and videotaped by 25 urologists and 51 residents in a simulated setting. The participants' degree of competence was assessed by a panel of eight independent expert urologists using the TOCO–TURBT tool. Each procedure was assessed by two raters. Feasibility, acceptability and content validity were evaluated by means of a quantitative cross-sectional survey. Regression analyses were performed to assess the

strength of the relation between experience and test scores (construct validity). Reliability was analysed by generalizability theory.

Results

The majority of assessors and urologists indicated the TOCO–TURBT tool to be a valid assessment of competency and would support the implementation of the TOCO–TURBT assessment as a certification method for residents. Construct validity was clearly established for all outcome measures of the procedural phase (all $r > 0.5$, $p < 0.01$). Generalizability-theory analysis showed high reliability (coefficient Phi ≥ 0.8) when using the format of two assessors and two cases.

Conclusions

This study provides first evidence that the TOCO–TURBT tool is a feasible, valid and reliable assessment tool for measuring competency in TURBT. The tool has the potential to be used for future certification of competencies for residents and urologists. The methodology of CTA might be valuable in the development of assessment tools in other areas of clinical practice.

Traditionally, declarations of competency in surgical skills have been based on the number of cases performed and the subjective opinion of a mentor, both indicating a perception of performance rather than an actual measurement of skills [1, 2]. The current shift from time-based residency training towards competency-based training has led to a growing demand for objective assessment of skills [3, 4]. Besides, several cases of technical incompetence have led to patient morbidity and mortality, which increased public and political pressure to evaluate surgical quality and competency [5, 6].

Objective assessment has the potential to measure competency before allowing independent clinical practice [1, 7]. Up till now, the majority of assessment tools have been used for formative assessment [8]. In formative assessment, the focus is on development and progression of residents throughout their traineeship by providing a structured evaluation of their performance and constructive feedback. In summative assessment, the goal is to assess the residents' performance at the end of a course by comparing it against some standard or benchmark. Summative assessment tools are used for high-stakes assessment and certification, which are imperative in competency-based training. For an assessment tool to be used for high-stakes assessment, it has to comply with stringent requirements, as it has to be objective, feasible, valid, and reliable [9].

In literature, the main focus in assessment has been on measurements of technical skills [1, 10]. However, a surgeon should be competent in non-technical skills as well [11]. Until now, few studies have focused on assessing both types of skills [11, 12]. 'Cognitive task analysis' (CTA) uses multiple interviews and observation strategies aiming at capturing experts' knowledge, thought processes, and decision-making on performance of a complex task [13, 14]. As experts have automated certain parts of their performance, they are no longer conscious of every step they take, which can lead to difficulties in the identification of decision points [13]. CTA offers a unique educational method to deconstruct the automated skills and identifies relevant steps

and decision points [15]. These steps and decision points are the basic elements of the assessment tool.
Transurethral resection of bladder tumour (TURBT) is a procedure every urologist should master. Being the initial bladder cancer treatment, competency in accurate and radical TURBT is paramount for achieving good staging and prognosis [16, 17]. Higher surgical experience results in less Ta-T1 cancer recurrence, and resident involvement in lower urinary tract surgery is associated with increased readmissions [18, 19]. This reflects the importance of training and assessments for minimizing risks for patients.
With this study, an assessment tool for TURBT is developed and validated. Research questions are "What are technical and non-technical skills necessary for a urologist to show competency in TURBT?" and "Is the newly developed Test Objective Competency (TOCO)–TURBT tool a valid and reliable assessment tool?"

SUBJECTS AND METHODS

Development of the TOCO–TURBT tool

The TOCO–TURBT tool was developed by means of a CTA. The stepwise method that was followed is reflected in Fig. 1. The first step was the design of a framework that consisted of a hierarchical description of all the constituent skills that enable competent performance of TURBT. It was designed by two urologists and two residents. Additional constituent skills were identified during 10 h of clinical observation. This resulted in a detailed overview of all discrete steps of the procedure, including a description of standards for acceptable performance and values for rating. An expert consensus meeting was organized, in which nine expert urologists participated. Urologists were selected based on their expertise in TURBT, familiarity with international guidelines, and involvement in the education of urological residents. Expert urologists were defined as registered oncological urologists with at least 10 years of practical experience, member of the European Association of Urology and the Dutch Association of Endourology and program director at their teaching hospital. During the consensus meeting, the content of the framework was extensively discussed until consensus on content, relevance, and completeness was reached. Subsequently, modifications to the framework were made according to the input and comments of the experts. Finally, the adjusted version of the assessment tool was sent to the expert panel for their final consent.

[FIGURE 1]
The final version of the TOCO–TURBT tool consists of 51 items and is divided into 3 phases: preparatory ($n = 15$), procedural phase ($n = 21$), and completion phase ($n = 15$) (Online Appendices 1a–c). The values for rating include yes/no and a 4-point Likert scale, in which scores three or four indicate competent performance of that particular item [20]. In addition, an assessment of overall performance (scale 1–10) and an expert global evaluation (competent/not competent) were included.

VALIDATION OF THE TOCO–TURBT TOOL

Study design and participants

This observational and comparative study was conducted at the urology departments of seven teaching hospitals across the Netherlands. A total of 76 residents and urologists with different levels of endoscopic experience were included. All participants received a standardized verbal introduction on the use of the validated Simbla TURBT simulator [21] from one of the tutors (HvG or HdV). Subsequently, each participant performed two standardized TURBT procedures. Standardization included the use of identical bladder substrates and the resection of four bladder tumours per procedure in a predefined order. Participants were instructed to perform a complete tumour resection until 2–3 mm below the surface of the bladder wall. The participants were blinded to the content of the TOCO–TURBT tool. No guidance, instruction or feedback was provided regarding the procedural steps nor regarding the technique of performance. To enable the assessment of cognitive skills (situational judgment, the participants' understanding of the procedure, salient decision points, etc.), a talk-aloud protocol was created, in which participants were instructed to express all their considerations and decisions throughout the procedure out loud [22]. Performance was recorded on video. Throughout each procedure, one of the tutors was present. A detailed description of the SIMBLA simulation model is provided in Online Appendix 2.

Video assessment

The video recordings were reviewed and scored with the TOCO–TURBT tool by a panel of eight independent expert urologists. To enhance the inter-rater reliability, the video assessments were preceded by an assessor meeting, in which four videos of participants with varying experience were reviewed and assessed. During this meeting, the description of standards for acceptable performance and the accompanying values for rating were clarified. The assessors received a handout of the standards per item to be used during the assessment.
The assessors were blinded for the participants' training status. Each procedure was assessed by a set of two raters. For practical reasons, the experts only assessed the procedural phase of the TURBT. The preparatory phase and the registration phase were assessed by two investigators (HdV and HvG), both medical doctors with an urological background. These investigators attended a similar assessor meeting prior to starting their assessments.

Assessment of feasibility, acceptability, and content validity

Data on demographics of participants, feasibility, acceptability, and content validity of the TOCO–TURBT tool were collected by means of a questionnaire, which was derived from Barton et al. [20]. The questionnaire was completed by the participants and the assessors. Question formats included multiple-choice questions and open-ended questions.

Assessment of construct validity

The following outcome measures were used to evaluate the construct validity of the TOCO–TURBT tool: (1) Test score per phase, subdivided into PrepScore (preparatory phase), ProcScore (procedural phase), and ComplScore (completion phase). (2) Overall performance (GlobalScore), (3) Competency score (ComptcyScore), and (4) Resection time (LogTime).

STATISTICAL ANALYSIS

The test score per phase was defined as the percentage credit points obtained over all items of that phase. Each item contributed a maximum of 1 point to the sum score. As there were scores from two assessors, the score was obtained by calculating the mean of the two scores. For each of the two cases, a proportion correct score was obtained by calculating the sum score over all the items and dividing it by the number of non-missing items. The final proportion correct score over cases was obtained by calculating the mean of proportion correct scores of cases 1 and 2.
For the procedural phase, mean scores were also obtained for the outcome measure 'overall performance' and 'competency score'. A skewed distribution was anticipated for the outcome measure 'resection time'. Therefore, a logarithmic transformation was applied to de-skew the distribution, and the resulting variable LogTime was used in the analysis. This variable was objectively measured by a single assessor (HdV).

RELIABILITY

The reliability of the TOCO–TURBT tool was estimated using generalizability theory (G-theory) [23]. With this approach, the variance of interest, between-participant differences in performance, is compared with the total variance in the performance measurements. The corresponding ratio is indicated as coefficient *Phi*, ranging between 0 and 1, and for high-stakes assessment values of 0.8 or higher are considered to indicate a sufficient level of reliability [24]. For technical details of this procedure, we refer to Online Appendix 3.

CONSTRUCT VALIDITY

For the assessment of construct validity, the relation between the outcome measures and the experience of participants was investigated. The experience of residents was weighted according to the levels of (in)dependent performance as recorded in their individual portfolio. For urologists, a distinction was made between partially and completely performed procedures. The equations used for the estimation of experience are described in Online Appendix 4.
The distribution of experience was extremely skewed and peaked (skewness = 2.8. kurtosis = 8.7), and was therefore transformed into a logarithmic scale. Regression analyses were performed to assess the strength of the relation between a participant's experience and the different outcome measures. The resulting regression coefficient (*b*) represents the slope of the regression line. The strength of the relation is indicated by the correlation coefficient (*r*) of the outcome measure and LogExp. According to Cohen's classification, correlations equal to 0.1, 0.3, and 0.5 correspond to small, moderate, and large effect sizes, respectively [25]. All statistical analyses were performed using SPSS version 22.

RESULTS

Between February and July 2015 a total of 76 participants were included in this study, 51 of whom were residents and 25 urologists. The general demographics of the participants and the assessors are described in Table 1.

FEASIBILITY, ACCEPTABILITY, AND CONTENT VALIDITY

[TABLE 1]

All assessors and participating urologists (strongly) agreed that the TOCO–TURBT tool covered all the important aspects of the TURBT procedure, and 93% of assessors and urologists (strongly) agreed that the tool seemed to be valid. All assessors considered the assessment process to be understandable and transparent, and indicated that the assessment tool corresponds with their professional judgement regarding competence. The majority of assessors and participants (88 and 62%, respectively) supported the implementation of the TOCO–TURBT assessment as a certification method before allowing residents at the end of their traineeship to independently perform a TURBT procedure on a patient. Moreover, the majority of assessors (88%) and participating urologists (95%) agreed that the TOCO–TURBT tool could be used for the certification of urologists.

RELIABILITY

The estimated variance components for the five outcome measures obtained in the G-analyses are shown in Online Appendix 5. For all measures, the largest component appeared to be the person variance (the variance of interest). This indicates that the person variance has the largest influence on the outcome measures, which is favourable for the reliability.

Table 2 shows the absolute reliability (coefficient *Phi*) resulting from the variance components presented in Online Appendix 5. As the table shows, the reliability is improved by an increase in the number of assessors and/or cases. In the current set-up of the TOCO–TURBT tool, two cases and two assessors were used for each assessment. With this set-up, satisfactory levels of reliability were obtained for all five measures (range 0.79–0.87).

[TABLE 2]

CONSTRUCT VALIDITY

Table 3 shows the descriptives and the results of the regression analyses for the six outcome measures, with LogExp as independent variable. The regression results for the preparation and completion phase (right panel of Table 3) showed no significant statistical relation between the participant's experience and the PrepScore and ComplScore. For the Procedural phase, all outcome measures showed a statistically significant linear relation with experience (LogExp). Moreover, the values of correlation (r) were in the range of 0.61–0.72, indicating large effect sizes. For LogTime, a negative relation with experience was found ($b = -0.076$), which is in agreement with the expectation that resection time in general will be lower for more experienced participants. The relation of this measure was not as strong as that of the other three procedural phase measures, but still substantial with a correlation (r) equal to − 0.34, which indicates a moderate effect size.

[TABLE 3]

DISCUSSION

This study clearly established the feasibility, content validity, and construct validity as well as the reliability of the TOCO–TURBT tool. This indicates that the TOCO–TURBT tool has the potential to be used for high-stakes assessment, such as certification of residents and relicensing of urologists.

The TOCO–TURBT tool was developed using the methodology of CTA, capturing automated expert knowledge that would otherwise have been lost [15]. Although this method is time-consuming and labour-intensive, it is recognized that a constructive approach to instrument development (including e.g. an a priori conceptual framework and use of expert consensus) helps to ensure that it captures the concept of interest, thereby increasing its content validity [26].

Feasibility and acceptability were confirmed, as the majority of assessors and participants indicated that they would support the implementation of the TOCO–TURBT assessment as a certification method for residents. For the assessment of construct validity, the broad classification of ‘novices’, ‘intermediates’, and ‘experts’ is generally used. This classification seems somewhat arbitrary and lacks precision. Therefore, we used experience as a continuous variable, derived from the validation study of the Program for Laparoscopic Urological Skills, performed by Tjiam et al. [27].

Construct validity was clearly demonstrated for the procedural phase of the TOCO–TURBT tool, but it was lacking for the preparatory and completion phase. The absence of construct validity in the preparatory and completion phase did not come as a surprise, and can be explained by the fact that the majority of items included in these phases are ‘checkpoints’, for which little experience or technical skill is needed. These two phases can be seen as the ‘basics’ of a TURBT procedure. The distinction between a novice and an expert becomes apparent in the procedural phase, where the actual TURBT is assessed, including the resection skills and handling complications. For certification purposes, the focus should be on this phase. The procedural phase of the TURBT procedures was assessed by a panel of expert urologists, whereas the other two phases were assessed by two general doctors. Ideally, the experts would have assessed the complete procedure, but unfortunately this was not possible due to time constraints. Still, we consider the current assessment approach to be valid, as the items of the preparation and completion phase do not need an ‘expert eye’ per se to be adequately assessed, whereas this is required for the procedural phase. The results of our study showed that the TOCO–TURBT tool is a reliable assessment tool, with two assessors and two trials being sufficient to reach substantial reliability.

To our knowledge, the TOCO–TURBT tool is the first assessment tool in the field of Urology that has the potential to be used for high-stakes assessment of TURBT in the future. With the current shift from time-based residency training towards competency-based residency training, objective assessment is becoming more important [3, 4]. In a recent project, a comparable tool that assesses basic laparoscopic skills has been developed and validated in The Netherlands [27]. In this project, the same methodology was used as in the current study. This “EBLUS” assessment is now a mandatory summative exam that junior residents have to pass

before they are allowed to participate in laparoscopic surgery. This example illustrates the possibilities of the future use of the TOCO–TURBT tool in the Dutch Urological curriculum.

Besides the potential use of the TOCO–TURBT tool for certification and relicensing, it could also be used in the light of the concept "Entrustable Professional Activities" (EPAs) [28]. An EPA is an activity residents can be trusted to perform competently in different stages of training. This concept translates competencies into clinical practice and enables supervisors to determine when a resident can be trusted to perform specific procedures with minimal supervision or without supervision. The TOCO–TURBT tool could be used to objectify a resident's performance throughout the different stages of independence. For this, the phases or components of the TURBT procedure that residents should master during different stages of their residency should be determined first.

The validation of the TOCO–TURBT tool was conducted in a simulation setting. The advantage of this approach is that the research setting was completely standardized, enabling a true comparison of performance without any interfering confounding factors. A drawback of this approach is that certain constituent skills, such as performing haemostasis and handling an obturator jurk, could not be simulated on the simulator and were only assessed in a cognitive way. This means that a participant that passes the exam knows the procedural steps, risks, and complications of the procedure, and competently performs the procedure on the simulation model. In 1990, Miller described a framework of increasingly complex levels of skills performance and assessment in the shape of a pyramid [29]. At the bottom of the pyramid there is knowledge (knows). On top of knowledge comes "knows how", "shows how", and finally "does". In our study, the assessment took place at the 'shows how' level. However, the unpredictability of the "does" level is not completely reproducible. Fortunately, literature has revealed that the measurement of the "shows how" correlates with the measurements of "does" [30].

We are planning to extrapolate the use of the TOCO–TURBT tool to the clinical setting. In this setting, we will have to make some adjustments to the tool. The simulation model could not mimic the exact layers of the bladder. Therefore, we were unable to assess the presence of muscularis propria in the specimen. In clinical setting, an item regarding the presence of muscularis propria in the specimen will be added. Still, it is important to realize that also workplace assessment has its limitations. When comparing with the drivers exam: a participant can show to master all the basic steps but can never encounter all possible dangerous traffic situations in an exam of 1 h.

Finally, the threshold of pass–fail is still subject of further study. The statistical method that we will use has previously been described by Tjiam et al. [31]. An important decision to make in the determination of a pass–fail standard is whether or not to use varying weights when calculating the total score from the item scores. It is important to realize that the different items are not independent from each other. A particular action/decision can be not very important in itself, but as it proceeds other actions/decisions it might have an important impact on the overall process. This makes the process of assigning weights quite arbitrary.

Up to now, we have studied the results for each phase (preparatory, procedural, and completion) separately. The results show a significant correlation between score and

experience for the procedural phase, whereas there was no significant correlation between score and experience for the preparation and completion phase. This may indicate that the procedural phase has the highest discriminatory power. In the decision of including weighing or not, perhaps the score of the procedural phase should have more weight when defining the pass–fail threshold.
If, after the pass/fail standard has been determined, a participant does not pass the exam, the consequence would be that he/she is not yet allowed to perform the procedure independently. More training is needed to ensure a safe situation and a new exam will be planned when the program director thinks the resident is ready for it.

CONCLUSION

This study provides first evidence that the TOCO–TURBT tool is a feasible, valid, and reliable high-stakes assessment tool for measuring technical and non-technical competency in TURBT. It has the potential to be used for the certification of skills of residents and urologists. The method of CTA might be of value in the development of assessment tools in other areas of clinical practice.

NOTES

Acknowledgements

The authors gratefully acknowledge all the experts who participated in the development and validation of the TOCO–TURBT tool, the residents and urologists who were willing to participate in this study, Ron Hoogeboom for statistical support, and Lisette van Hulst for editorial assistance.

Compliance with ethical standards

Disclosures

A.H. de Vries, H.G.J van Genugten, A.J.M. Hendrikx, E.L. Koldewijn, B.M.A. Schout, A.M.M. Muijtjens, C.P.M. van der Vleuten, C. Wagner, Irene M. Tjiam, and J.J.G. van Merriënboer have no conflicts of interest or financial ties to disclose.

Ethical approval

Ethical approval was sought from the institution's research and ethics committee. Ethical approval was not required according to the Dutch Medical Research (Human Subjects) Act, since no patients or patient details were involved.

Informed consent

Informed consent with assurance of anonymity and confidentiality was obtained from all participants.

REFERENCES

1. Aggarwal R, Grantcharov T, Moorthy K, Milland T, Darzi A (2008) Toward feasible, valid, and reliable video-based assessments of technical surgical skills in the operating room. Ann Surg 247:372–379. https://doi.org/10.1097/SLA.0b013 e318160b37 1
2. Zevin B, Bonrath EM, Aggarwal R, Dedy NJ, Ahmed N, Grantcharov TP, ATLAS group (2013) Development, feasibility, validity, and reliability of a scale for objective assessment of operative performance in laparoscopic gastric bypass surgery. J Am Coll Surg 216:955–965.e8. https ://doi.org/10.1016/j.jamco llsurg.2013.01.003 (**quiz 1029-31, 1033**)
3. Hampton T (2015) Efforts seek to develop systematic ways to objectively assess surgeons' skills. JAMA 313:782–784. https ://doi.org/10.1001/jama.2015.233
4. de Montbrun S, Satterthwaite L, Grantcharov TP (2016) Setting pass scores for assessment of technical performance by surgical trainees. Br J Surg 103:300–306. https ://doi.org/10.1002/bjs.10047
5. Baker GR, Norton PG, Flintoft V, Blais R, Brown A, Cox J, Etchells E, Ghali WA, Hebert P, Majumdar SR, O'Beirne M, Palacios-Derflingher L, Reid RJ, Sheps S, Tamblyn R (2004) The Canadian Adverse Events Study: the incidence of adverse events among hospital patients in Canada. CMAJ 170:1678–1686
6. Vincent C, Moorthy K, Sarker SK, Chang A, Darzi AW (2004) Systems approaches to surgical quality and safety: from concept to measurement. Ann Surg 239:475–482
7. Miskovic D, Ni M, Wyles SM, Kennedy RH, Francis NK, Parvaiz A, Cunningham C, Rockall TA, Gudgeon AM, Coleman MG, Hanna GB, National Training Programme in Laparoscopic Colorectal Surgery in England (2013) Is competency assessment at the specialist level achievable? A study for the national training programme in laparoscopic colorectal surgery in England. Ann Surg 257:476–482. https ://doi.org/10.1097/SLA.0b013 e3182 75b72 a
8. Fried GM, Feldman LS (2008) Objective assessment of technical performance. World J Surg 32:156–160. https ://doi.org/10.1007/s0026 8-007-9143-y
9. Gallagher AG, Ritter EM, Satava RM (2003) Fundamental principles of validation, and reliability: rigorous science for the assessment of surgical education and training. Surg Endosc 17:1525–1529. https ://doi.org/10.1007/s0046 4-003-0035-4
10. Larson JL, Williams RG, Ketchum J, Boehler ML, Dunnington GL (2005) Feasibility, reliability and validity of an operative performance rating system for evaluating surgery residents. Surgery 138:640–647 (**discussion 647–9**)
11. Hull L, Arora S, Aggarwal R, Darzi A, Vincent C, Sevdalis N (2012) The impact of nontechnical skills on technical performance in surgery: a systematic review. J Am Coll Surg 214:214–230.https ://doi.org/10.1016/j.jamco llsur g.2011.10.016
12. Shepherd W, Arora KS, Abboudi H, Shamim Khan M, Dasgupta P, Ahmed K (2014) A review of the available urology skills training curricula and their validation. J Surg Educ 71:289–296. https://doi.org/10.1016/j.jsurg .2013.09.005
13. Clark R, Feldon D, van Merrienboer J, Yates K (2008) Cognitive task analysis. Anonymous Handbook of research on educational communications and technology. Taylor and Francis Group, Boca Raton, pp 577–593
14. Yates K, Feldon D (2011) Advancing the practice of cognitive task analysis: a call for taxonomic research. Theor Issues Ergon Sci 11:1464–1536
15. Sullivan ME, Ortega A, Wasserberg N, Kaufman H, Nyquist J, Clark R (2008) Assessing the teaching of procedural skills: can cognitive task analysis add to our traditional teaching methods? Am J Surg 195:20–23
16. Babjuk M, Burger M, Zigeuner R, Shariat SF, van Rhijn BW, Comperat E, Sylvester RJ, Kaasinen E, Bohle A, Palou Redorta J, Roupret M, European Association of Urology (2013) EAU guidelines on non-muscle-invasive urothelial carcinoma of the bladder: update 2013. Eur Urol 64:639–653. https ://doi.org/10.1016/j.eururo.2013.06.003
17. Brausi M, Collette L, Kurth K, van der Meijden AP, Oosterlinck W, Witjes JA, Newling D, Bouffioux C, Sylvester RJ, EORTC Genito-Urinary Tract Cancer Collaborative Group (2002) Variability in the recurrence rate at first follow-up cystoscopy after TUR in stage Ta T1 transitional cell carcinoma of the bladder: a combined analysis of seven EORTC studies. Eur Urol 41:523–531

18. Allard CB, Meyer CP, Gandaglia G, Chang SL, Chun FK, Gelpi-Hammerschmidt F, Hanske J, Kibel AS, Preston MA, Trinh QD (2015) The effect of resident involvement on perioperative outcomes in transurethral urologic surgeries. J Surg Educ 72:1018–1025
19. Jancke G, Rosell J, Jahnson S (2014) Impact of surgical experience on recurrence and progression after transurethral resection of bladder tumour in non-muscle-invasive bladder cancer. Scand J Urol 48:276–283. https://doi.org/10.3109/21681 805.2013.864327
20. Barton JR, Corbett S, van der Vleuten CP, English Bowel Cancer Screening Programme, UK Joint Advisory Group for Gastrointestinal Endoscopy (2012) The validity and reliability of a Direct Observation of Procedural Skills assessment tool: assessing colonoscopic skills of senior endoscopists. Gastrointest Endosc 75:591–597. https://doi.org/10.1016/j.gie.2011.09.053
21. de Vries AH, van Genugten HG, Hendrikx AJ, Koldewijn EL, Schout BM, Tjiam IM, van Merrienboer JJ, Muijtjens AM, Wagner C (2016) The Simbla TURBT simulator in urological residency training: from needs analysis to validation. J Endourol. https://doi.org/10.1089/end.2015.0723
22. Luker KR, Sullivan ME, Peyre SE, Sherman R, Grunwald T (2008) The use of a cognitive task analysis-based multimedia program to teach surgical decision making in flexor tendon repair. Am J Surg 195:11–15
23. Crossley J, Davies H, Humphris G, Jolly B (2002) Generalisability: a key to unlock professional assessment. Med Educ 36:972–978
24. Fraenkel JR, Wallen NE (2009) How to design and evaluate research in education. McGraw Hill, Boston
25. Cohen J (1988) Statistical power analysis for the behavioural sciences. Lawrence Erlbaum, London
26. Walsh CM, Ling SC, Khanna N, Cooper MA, Grover SC, May G, Walters TD, Rabeneck L, Reznick R, Carnahan H (2014) Gastrointestinal endoscopy competency assessment tool: development of a procedure-specific assessment tool for colonoscopy. Gastrointest Endosc 79(e5):798–807. https ://doi.org/10.1016/j.gie.2013.10.035
27. Tjiam IM, Persoon MC, Hendrikx AJ, Muijtjens AM, Witjes JA, Scherpbier AJ (2012) Program for laparoscopic urologic skills: a newly developed and validated educational program. Urology 79:815–820. https ://doi.org/10.1016/j.urolo gy.2012.01.014
28. Ten Cate O (2013) Nuts and bolts of entrustable professional activities. J Grad Med Educ 5:157–158. https ://doi.org/10.4300/JGME-D-12-00380 .1
29. Miller GE (1990) The assessment of clinical skills/competence/performance. Acad Med 65:S63-7
30. Ram P, van der Vleuten C, Rethans JJ, Grol R, Aretz K (1999) Assessment of practicing family physicians: comparison of observation in a multiple-station examination using standardized patients with observation of consultations in daily practice. Acad Med 74:62–69
31. Tjiam IM, Schout BM, Hendrikx AJ, Muijtjens AM, Scherpbier AJ, Witjes JA, Van Der Vleuten CP (2013) Program for laparoscopic urological skills assessment: setting certification standards for residents. Minim Invasive Ther Allied Technol 22:26–32. https://doi.org/10.3109/13645 706.2012.68691 8

FIGURES AND TABLES

Fig. 1 : Stepwise method for the development of the TOCO–TURBT tool

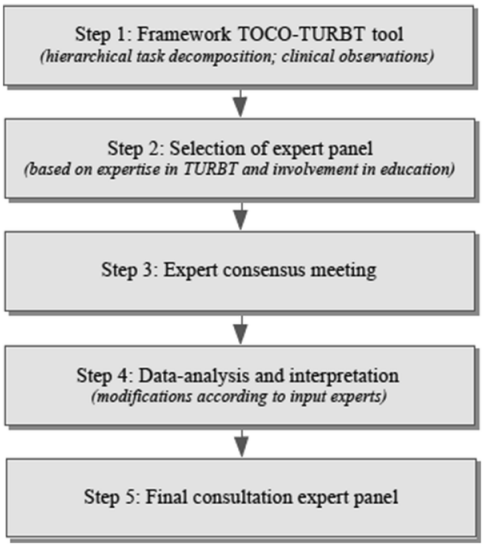

Table 1 General demographics

	Age (years)	Gender (m/f)	Dexterity (right/left)	Experience in TURBT (independently performed)
PG Y0 (n = 8)	26.5 (25–29)	3/5	6/2	0 (0)
PG Y1 (n = 3)	28 (28)	2/1	3/0	0 (0)
PG Y2 (n = 4)	29.5 (28–30)	2/2	3/1	0 (0)
PG Y3 (n = 12)	30 (28–34)	8/4	12/0	0 (0–20)
PG Y4 (n = 6)	30 (30–36)	1/5	5/1	0 (0–1)
PG Y5 (n = 11)	32 (31–38)	8/3	9/2	6 (0–30)
PG Y6 (n = 7)	34 (32–37)	2/5	7/0	6 (2–38)
Urologist (n = 25)	49 (32–64)	23/3	24/1	100 (50–500)
Assessor (n = 8)	54 (43–68)	8/0	8/0	350 (100–1000)

Values for variables (age and experience in TURBT) are presented as medians with range (min–max)

PG Y0 post graduate student, not yet in training, *PG Y1–6* resident in training year 1–6

Table 2 : Reliability (coefficient Phi) of the outcome measures in relation to the numbers of assessors (N_a) and cases (N_c) used to assess a participant with the TOCO–TURBT tool

	Preparatory phase			Procedural phase									Completion phase		
	PrepScore			ProcScore			GlobalScore			ComptcyScore			ComplScore		
N_a	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
N_c															
1	0.74	0.79	0.81	0.61	0.73	0.78	0.65	0.77	0.83	0.51	0.65	0.71	0.62	0.66	0.67
2	0.83	**0.87**	0.89	0.76	**0.84**	0.87	0.79	**0.87**	0.91	0.68	**0.79**	0.83	0.76	**0.79**	0.80
3	0.86	0.90	0.91	0.83	0.89	0.91	0.85	0.91	0.93	0.76	0.85	0.88	0.82	0.85	0.86

As we focused on 2 cases and 2 assessors these numbers were made bold
For high-stakes assessment a *Phi* of 0.8 or higher is generally considered sufficient [24].

Table 3 Descriptives and regression analysis results for the outcome measures with the logarithm of experience as independent variable

Phase	Performance measure	Descriptives[a]					Regression with LogExp[b]				
		M	SD	Min	Max	**N**	**b**	**r**	**p**	95% CI	
										lo	hi
Preparatory	PrepScore (0–100)	63	11	27	88	76	2.3	0.17	0.148	− 0.8	5.4
Procedural	ProcScore (0–100)	65	17	10	91	76	12	0.61	0.001	9	16
	GlobalScore (1–10)	6.3	1.7	2.0	8.5	76	1.5	0.72	0.001	1.2	1.8
	ComptcyScore (0–1)	0.56	0.41	0.00	1.00	76	0.35	0.72	0.001	0.27	0.43
	LogTime (min)	1.91	0.19	1.30	2.38	75	− 0.076	− 0.34	0.003	− 0.13	− 0.03
Completion	ComplScore (0–100)	60	12	31	88	76	1.4	0.10	0.395	− 1.8	4.6

[a]*M* mean, *SD* standard deviation, *min/max* minimum/maximum, *N* number of subjects with non-missing value
[b]*b*:regression coefficient, *r* correlation, *p p* value, *lo-hi* lower and higher boundary of the 95% CI of *b*