Mastoidectomy performance assessment of virtual simulation training using final-product analysis

Steven A. W. Andersen, MD; Per Cayé-Thomasen, MD, DMSc; Mads Sølvsten Sørensen, MD, DMSc

Department of Otorhinolaryngology – Head & Neck Surgery, University Hospital of Copenhagen/Rigshospitalet, Denmark

Conflict of interests:

None.

Financial disclosures:

The development of the Visible Ear Simulator software has been financially supported by the Oticon Foundation.

Full citation: Andersen SA, Cayé-Thomasen P, Sølvsten Sørensen M. Mastoidectomy Performance Assessment of Virtual Simulation Training Using Final-Product Analysis. Laryngoscope. 2015 Feb; 125(2):431–5.

DOI: 10.1002/lary.24838

Objectives: The future development of integrated automatic assessment in temporal bone virtual surgical simulators calls for validation against currently established assessment tools. This study aims to explore the relationship between mastoidectomy final-product performance assessment in virtual simulation and traditional dissection training.

Study Design: Prospective trial with blinding.

Methods: A total of 34 novice residents performed a mastoidectomy on the Visible Ear Simulator and on a cadaveric temporal bone. 2 blinded, senior otologists assessed the final-product performance using a modified Welling Scale. The simulator gathered basic metrics on time, steps and volumes in relation to the on-screen tutorial and collisions with vital structures.

Results: Substantial inter-rater reliability (kappa=0.77) for virtual simulation and moderate inter-rater reliability (kappa=0.59) for dissection final-product assessment was found. The simulation and dissection performance scores had significant correlation (p=0.014). None of the basic simulation metrics correlated significantly with final-product score except for number of steps completed in the simulator.

Conclusion: A modified version of a validated final-product performance assessment tool can be used to assess mastoidectomy on virtual temporal bones. Performance assessment of virtual mastoidectomy could potentially save the use of cadaveric temporal bones for more advanced training when a basic level of competency in simulation has been achieved.

Key-words: virtual reality simulation, surgical simulation training, temporal bone dissection, final-product assessment, mastoidectomy training

Level of evidence: IIb

INTRODUCTION

Temporal bone skills are considered key for the otorhinolaryngology (ORL) resident in training and have traditionally been taught to novices through cadaveric temporal bone dissection. The assessment of technical competency in mastoidectomy training can be based on expert opinion or one of the more structured tools that have been developed. These instruments capture different aspects of the procedure for example technical skills, process and final product. A global rating scale (GRS), a task-based checklist (TBC) and final-product analysis (FPA) was proposed by a group from Toronto[1] and a similar approach consisting of a checklist of procedural steps and a 10-item global preparation and process scale originates from Johns Hopkins[2]. The Welling Scale (WS1) as reported by Butler et al consists of a 35-item binary grading instrument for final-product analysis[3]. The performance assessment of technical competency can for example be used for formative feedback and monitoring of skills development of the resident[2].

There is an increasing evidence-base for virtual simulation training in temporal bone surgery and virtual simulation allows for repeat training and multiple assessments in a standardized and reproducible environment. Simulator metrics have been shown for example to be able to discriminate between novices and experts[4-8] and Wiet et al. have proposed the use of a unifying cross-institutional assessment scale[9] for the structured assessment of novices in virtual temporal bone surgical simulation[10]. Automated and objective assessment in the virtual surgical simulators seems possible with the development of more advanced metrics for the measurement of final-product-like items[11]. The development of future simulator-integrated and valid automatic assessment using simple and advanced simulator metrics calls for the validation of simulator assessment using already established assessment tools and a standard setting score will need to be defined.

To date no study has explored whether the currently validated final-product assessment tools for competency in mastoidectomy can be used to assess a virtual performance, which would provide further content validity evidence for the scale. The hypothesis is that final-product analysis can be used for the assessment of virtual mastoidectomy performance. In this study we used a modified Welling Scale for comparing virtual simulation and dissection final-product performance.

MATERIAL AND METHODS

A total of 34 ORL residents participated in the national temporal bone course at our department in January 2012 (17 participants) and 2013 (17 participants) and were included in this study. The residents were post-graduate year 2-5 and were all novices with no hands-on experience in temporal bone surgery and only limited virtual temporal bone simulation experience. The participants performed a complete mastoidectomy with entry into the antrum on a virtual temporal bone in the Visible Ear Simulator version 1.2. The simulator provided basic mastoidectomy tutoring with volumetric green lighting of the volume to be drilled in each step along with a step-by-step onscreen tutorial corresponding to a classical temporal bone dissection manual. The participants were teamed in pairs during simulation and had to divide 80 minutes of virtual training between them. The virtual temporal bone and simulator metrics were saved for later final-product assessment. One participant had no virtual bone for assessment due to a computer crash and was excluded. The following day the participants performed a complete mastoidectomy with entry into the antrum on a cadaveric temporal bone. The participants were teamed as in simulation and were each allowed about 60 minutes for the procedure using a temporal bone dissection manual and feedback by four senior otologists. The participants thereby had feedback and guidance in both training modalities.

The Visible Ear Simulator (VES) version 1.2 is a fully functional 3D virtual temporal bone simulator with force-feedback[12, 13]. The simulator was developed by the senior author (MSS) in collaboration with Peter Trier Mikkelsen of the Alexandra Institute and is freeware and available for download from the group's homepage[14]. The simulator features a single temporal bone and runs on a standard PC with a GeForce GTX® graphics card and a Phantom Omni® haptic device (now Geomagic® Touch™) for force-feedback and intuitive drill handling or a mouse (no force-feedback). VES also features an integrated tutor-function and a step-by-step tutorial for mastoidectomy and is available in multiple languages. The simulator gathers basic metrics on the time used for the procedure, the number of steps completed in the tutorial, the amount of bone removed inside and outside of the reference volume and the collisions with the dura, the middle ear bones, the inner ear and the facial nerve.

Two senior otologists (PCT and MSS) assessed the virtual and dissected temporal bones using a modified Welling Scale with 25 binary items for final-product analysis (appendix 1). The raters were blinded to which virtual and cadaveric temporal bone the participants had drilled. The original scale was modified slightly to reflect the procedural steps in our setting. The Visible Ear Simulator allows for saved virtual temporal bones to be opened and examined with all degrees of freedom. In rating the virtual temporal bones the raters used the simulator metrics on collisions to assess whether vital structures were identified, exposed and untouched because the vital structures are programmed as reference structures in the virtual model and collisions therefore leaves no visual trace as it otherwise would on a real temporal bone.

The collected data were analyzed with SPSS (SPSS Inc., Chicago, IL) version 20 for MacOS X using linear regression, ANOVA and inter-rater reliability kappa-statistics.

RESULTS

An example of the final-product of a dissected and virtual temporal bone is shown in figure 1. The performance score assigned by the two raters showed highly significant linear correlation in both simulation (p<0.001) and dissection (p<0.001) (Figure 2A and 2B). This corresponds to a substantial inter-rater reliability with a kappa of 0.77 (95% CI [0.72-0.81]) for the simulation performance score and moderate inter-rater reliability kappa of 0.59 (95% CI [0.54-0.64]) for the dissection performance score.

The simulation and dissection performance final-product scores had significant linear correlation (p=0.014) (Figure 3).

The association between simulator metrics and the final-product performance score in simulation metrics were analyzed for correlation using linear regression (Table 1). A significant correlation (p<0.001) between the simulator performance score and the number of tutorial steps completed was found. For the drilled volume in percent of the reference volume a non-significant trend (p=0.061) towards a higher performance score with more of the reference volume removed could be demonstrated. For time, volume outside reference volume and number of collisions, no correlation between the metric and the simulation final-product score was found.

No effect of the order in the teams was found in simulation or dissection using ANOVA meaning that being the partner performing the procedure first or last had no influence on the final-product score (in simulation mean 14.88 and 14.91, respectively (p=0.98); in dissection mean 13.72 and 12.50 respectively (p=0.31)).

DISCUSSION

The inter-rater reliability kappa coefficient of the modified Welling Scale was determined in both virtual and dissection mastoidectomy performance and substantial agreement on the virtual mastoidectomy and moderate agreement on the dissection mastoidectomy was found. Butler and Wiet reported a moderate inter-rater agreement of kappa=0.49 – 0.64 for the original Welling Scale (WS1) using 6 raters who on two occasions rated both cadaveric and plastic temporal bones[3]. Zhao et al also used a modified Welling Scale for the assessment of temporal bone performance with two raters and found a moderate inter-rater reliability of the scale with a kappa coefficient of 0.47[15]. The inter-rater reliability kappa coefficient of 0.59 found in our study for the dissected bones is comparable to previous reports. Our modifications to the original WS1 were done to reflect the procedure in our setting, and we have no reason to believe that these minor changes would have made the tool less valid or have any impact on the validity of the scale, as the design principles behind the scale were unchanged.

A significantly higher and substantial inter-rater reliability kappa was found for the virtual temporal bones. Possible explanations for this finding could be that the virtual bone was the same for all the participants, that the raters were familiar with this bone and also that the simulator metrics on collision were used as a supplement in determining whether vital structures were identified, exposed and untouched. Using a virtual system for measuring performance thereby offers some advantages over cadaveric dissection in final-product assessment. If valid and reliable metrics can be integrated in the simulator for automatic assessment then a truly objective measure of performance could be used for both formal assessment and for training and skills development with simulator feedback.

Final-product assessment tools consist of checklists of items defined by experts as criteria for a good performance. Each item of the Welling Scale is binary (adequate/inadequate) and a maximum score would be the expected competency level of all experts. A maximum or near-maximum score would also be the benchmark level of resident performance competency in order to operate supervised. While analysis of final-product performance with an assessment tool like the WS is very valuable in monitoring the development of skills at the residency level of training the inherent limitations of binary scoring of each item makes the scale inadequate for more complex assessment and score standardization. The final-product has the advantage that it can be saved for later analysis but final-product assessment also has limitations as it considers only the result and not directly the process or the technical competency. Final-product items are however found to be relevant in the assessment of mastoidectomy performance[9] and could be integrated in future simulator-based assessment[11].

Some of the other standardized, objective assessment tools developed in Toronto and at Johns Hopkins use Likert-scaled items that provide a more continuous scale for grading and a valid pass/failure benchmark performance for necessary competency level in mastoidectomy could be set for these types of assessment tools using the contrasting groups method (discrimination between novices and experts). These tools however require direct observation of performance[1, 2], which would be more time and resource consuming as either additional trained assessors are needed for rating or fewer trainees could be assessed at the same time.

In a subsequent analysis using generalization theory Fernandez, Butler, Wiet et al found that WS1 performance ratings were consistent across raters and rating sessions and that it was residents' inconsistent performance across different bones that introduced most of the measurement variation[16]. They conclude that performance assessment using the WS1 should be based on the evaluation of multiple performances rather than introducing more raters. Even though we found a significantly higher inter-rater reliability for the modified WS in virtual simulation in this study we believe that assessment of multiple virtual performances should be done to ensure reliability. This is especially important if assessment is to be used for other purposes than formative feedback and monitoring of skills development for example in high-stakes pass/fail assessment. With the future integration of valid automated assessment into virtual surgical simulators assessment of multiple procedures seems feasible.

The inconsistent performance by the same participant between procedures found by Fernandez et al could also explain some of the variance seen in figure 3. In spite of this a significant correlation between the simulator and dissection final-product performance was found. This could indicate that at the novice level of training some of the same competencies and skills may be acquired in the virtual simulation setting as in the dissection setting. The virtual simulation training which has lower fidelity than dissection could then be used for the initial training of residents to conserve scarce educational resources for more advanced training once basic competencies have been achieved.

In the described temporal bone course under which the study was conducted, the focus was to let participants have the maximal outcome from a single temporal bone dissection session without intervening in the current course setup. Therefore a variety of factors could affect our results by tending to cancel out differences in the skill levels of the participants. In both virtual and dissection training the participants were teamed together in pairs and a observer-first and/or peer feedback, as well as unequal division of time between participants could have influenced final-product performance score. The order of the team participants could however not be demonstrated to have a statistically significant effect in our analyses. In the simulator, the participants were guided by green lighting of the intended volume to be drilled along with a step-by-step tutorial. In cadaveric dissection the participants had constant feedback from four experienced senior instructors as well as several plenary sessions with feedback. The participants had more time in dissection training, but also had to use more complex psychomotor skills with the suction device, which is not incorporated into the simulation environment. In future studies, these factors should be controlled in order to explore the relationship between simulator and dissection performance further.

None of the current simulator metrics except for number of steps performed were found to be statistically correlated with the final-product performance score. The WS consist of groups of items ordered in the progressive steps of a mastoidectomy and therefore the number of completed steps should directly reflect on the final-product performance score. Our findings are consistent with this. A trend towards association between reference volume (which is calculated dependent of the current step in the mastoidectomy) and the final-product score was found. This association would be expected, as many of the final-product scale items are related to volume like for example exposure and ‘no remaining cells’. None of the basic metrics in the Visible Ear Simulator can thereby currently be used for automatic assessment.

CONCLUSION

In developing and validating simulator-based assessment, the comparison with performance scores on an already validated performance assessment tool is necessary. It was demonstrated that a modified Welling Scale could be used to assess virtual temporal bones with a substantial degree of inter-rater reliability. A significant relationship between the final-product performance in simulation and dissection was demonstrated. This could point towards the simulator being a relevant training tool for dissection at the novice level of training as performance on the two modalities is correlated. A virtual system can thereby be used for formative feedback and monitoring of skills development. This could save the use of cadaveric temporal bones and human instructors for the next level of training, when basic competencies have been achieved in simulation. Currently, there is still much work needed in developing new and valid simulator metrics for the objective, structured assessment of mastoidectomy performance and for defining a benchmark performance for necessary competency level in mastoidectomy, both in simulation and in traditional training.

ACKNOWLEDGMENTS

The authors would like to acknowledge Peter Trier Mikkelsen and Karsten Østergaard Noe of the Alexandra Institute, Aarhus, for their continuing contribution to the development of the Visible Ear Simulator. The development of the Visible Ear Simulator has been financially supported by the Oticon Foundation.

REFERENCES

[1] Zirkle M, Taplin MA, Anthony R & Dubrowski A. Objective assessment of temporal bone drilling skills. Ann Otol Rhinol Laryngol 2007; 116:793-798.

[2] Laeeq K, Bhatti NI, Carey JP et al. Pilot testing of an assessment tool for competency in mastoidectomy. Laryngoscope 2009; 119:2402-2410.

[3] Butler NN & Wiet GJ. Reliability of the Welling scale (WS1) for rating temporal bone dissection performance. Laryngoscope 2007; 117:1803-1808.

[4] Sewell C, Morris D, Blevins NH et al. Validating metrics for a mastoidectomy simulator. Stud Health Technol Inform 2007; 125:421-426.

[5] Sewell C, Morris D, Blevins NH et al. Providing metrics and performance feedback in a surgical simulator. Comput Aided Surg 2007; 13:63-81.

[6] Zirkle M, Roberson DW, Leuwer R & Dubrowski A. Using a virtual reality temporal bone simulator to assess otolaryngology trainees. Laryngoscope 2007; 117:258-263.

[7] Khemani S, Arora A, Singh A et al. Objective skills assessment and construct validation of a virtual reality temporal bone simulator. Otol Neurotol 2012; 33:1225-1231.

[8] Zhao YC, Kennedy G, Hall R & O'Leary S. Differentiating levels of surgical experience on a virtual reality temporal bone simulator. Otolaryngol Head Neck Surg 2010; 143(5 Suppl 3):30-35.

[9] Wan D, Wiet GJ, Welling DB et al. Creating a cross-institutional grading scale for temporal bone dissection. Laryngoscope 2010; 120:1422-1427.

[10] Wiet G, Hittle B, Kerwin T & Stredney D. Translating surgical metrics into automated assessments. Stud Health Technol Inform 2012; 173:543-548.

[11] Kerwin T, Wiet G, Stredney D & Shen HW. Automatic scoring of virtual mastoidectomies using expert examples. Int J Comput Assist Radiol Surg 2012; 7:1-11.

[12] Trier P, Noe KO, Sorensen MS & Mosegaard J. The visible ear surgery simulator. Stud Health Technol Inform 2008; 132:523-525.

[13] Sorensen MS, Mosegaard J & Trier P. The visible ear simulator: a public PC application for GPU-accelerated haptic 3D simulation of ear surgery based on the visible ear data. Otol Neurotol 2009; 30:484-487.

[14] The Visible Ear Project group Web site. http://ves.cg.alexandra.dk/. Accessed January 27. 2014.

[15] Zhao YC, Kennedy G, Yukawa K et al. Can virtual reality simulator be used as a training aid to improve cadaver temporal bone dissection? Results of a randomized blinded control trial- Laryngoscope 2011; 121:831-837.

[16] Fernandez SA, Wiet GJ, Butler NN et al. Reliability of surgical skills scores in otolaryngology residents: analysis using generalizability theory. Eval Health Prof 2011; 31:419-436.

Table 1. Simulator metrics and correlation with final-product performance score on the modified Welling Scale.

Simulator metric	Mean (range)	Correlation with final-product performance score
Total time	38.1 min (20.2-59.0 min)	non-significant, p=0.542)
Steps in the on-screen tutorial manual completed	9.1 (5-14)	significant, p<0.001)
Volume inside reference	82.1 % (58.1-97.3 %)	non-significant, p=0.061
Volume outside reference	1.4 ccm (0.1-5.72 ccm)	non-significant, p=0.387
Total collisions with vital structures	8.2 (0-106)	non-significant , p=0.192

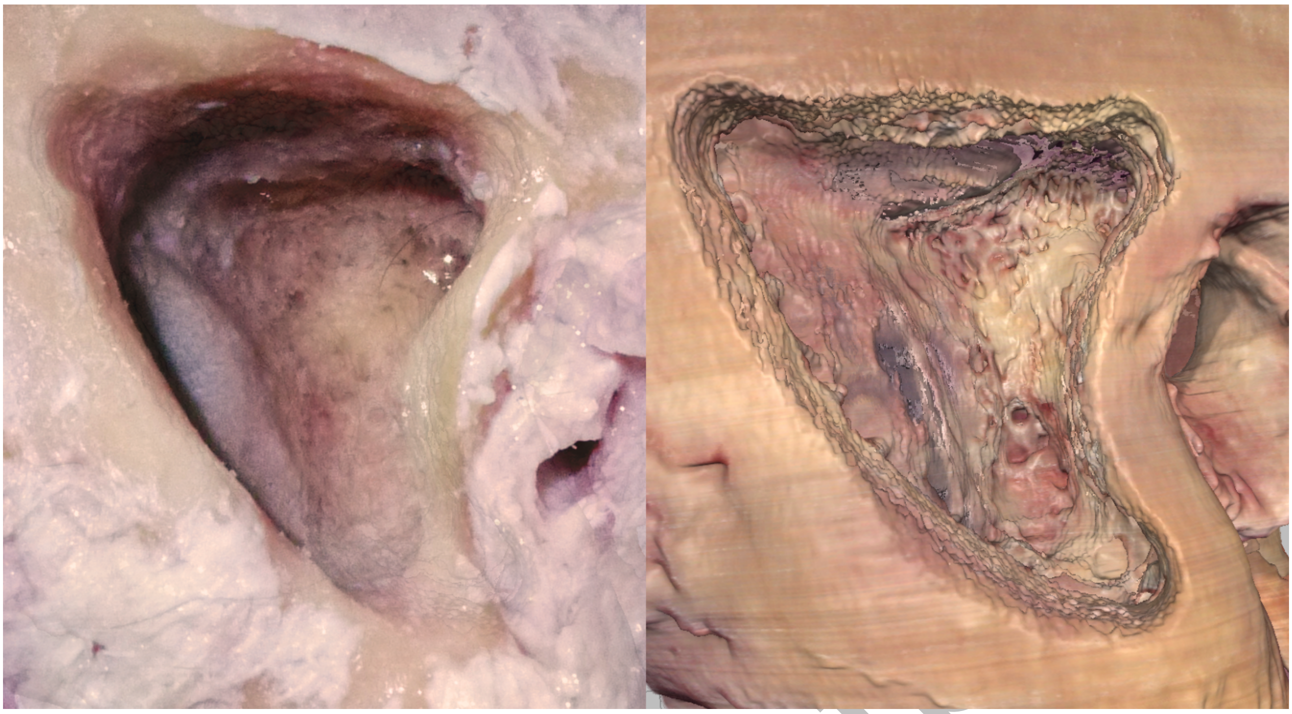

Figure 1. An example of a dissection mastoidectomy (left) and a comparable virtual mastoidectomy in the Visible Ear Simulator (right).

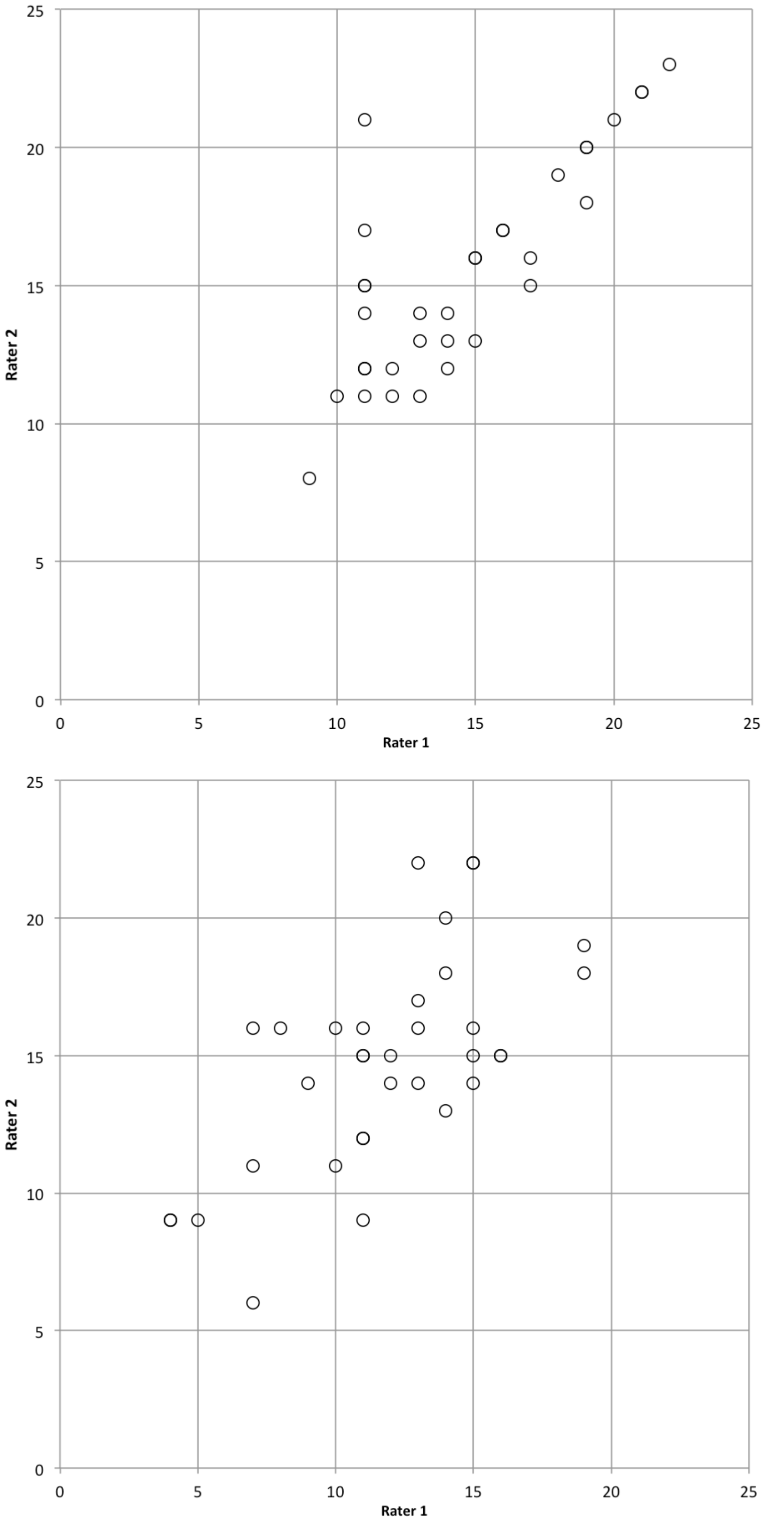

Figure 2. A) Simulator final-product score and B) Dissection final-product score, assigned by the two expert raters.

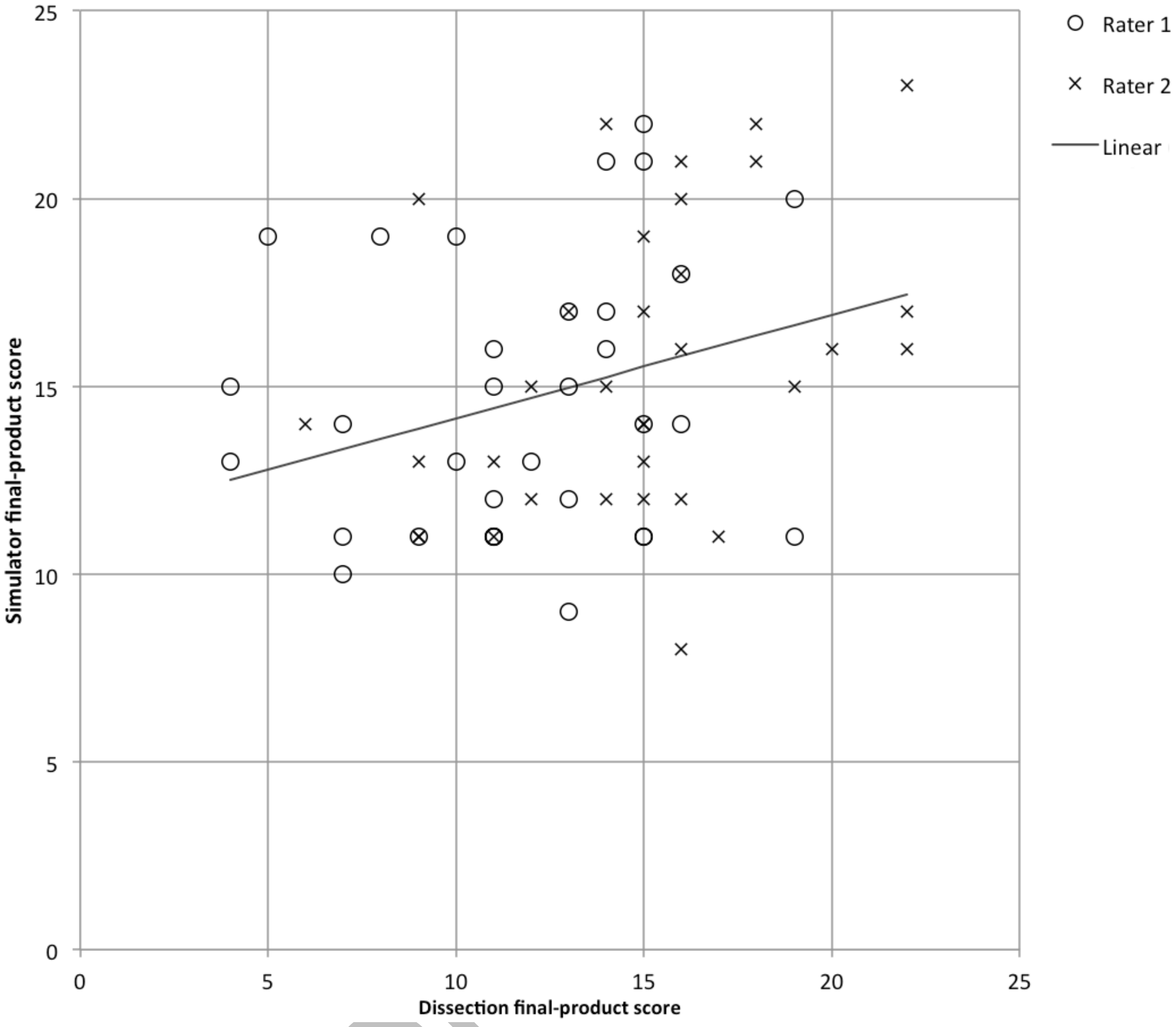

Figure 3. The correlation between simulator and dissection final-product performance scores assigned by the two expert raters.

Temporal Bone Dissection Outcome Performance - Modified Welling Scale

Grade each item: 0 = incomplete/inadequate dissection, 1 = complete

Mastoidectomy margins defined at:		
1. Temporal line	0	1
2. Posterior canal wall	0	1
3. Sigmoid sinus	0	1
Antrum mastoideum		
4. Antrum entered	0	1
5. Lateral semicircular canal exposed	0	1
6. Lateral semicircular canal intact	0	1
Sigmoid sinus		
7. Exposed, no overhang	0	1
8. No cells remain	0	1
9. No holes	0	1
Sinodural angle		
10. Sharp	0	1
11. No cells remain	0	1
Tegmen mastoideum/tympani		
12. Attic/tegmen tympany exposed	0	1
13. Ossicles intact (untouched)	0	1
14. Tegmen mastoideum exposed	0	1
15. No cells remain	0	1
16. No holes	0	1
Mastoid tip		
17. Digastric ridge exposed	0	1
18. Digastric ridge followed towards stylomastoid foramen	0	1
19. No cells remain	0	1
External auditory canal		
20. Thinning of the posterior canal wall	0	1
21. No cells remain	0	1
22. No holes	0	1
Facial nerve		
23. Facial nerve identified (vertical part)	0	1
24. No exposed nerve sheath	0	1
25. Tympanic chorda exposed	0	1

Appendix 1. The modified Welling Scale.