Manuscript title:

Re-defining the virtual reality dental simulator: Demonstrating concurrent validity of clinically relevant assessment and feedback

Running title:

Re-defining the virtual reality dental simulator

Dixon, Jonathan[1]
Jonathan.dixon@sheffield.ac.uk

Towers, Ashley[1]
a.towers@sheffield.ac.uk

Martin, Nicolas[1]
n.martin@sheffield.ac.uk

Field, James[2]
Fieldj2@cardiff.ac.uk

1 Academic Unit of Restorative Dentistry
School of Clinical Dentistry
The University of Sheffield
UK

2 School of Dentistry
Cardiff University
UK

Acknowledgements:

The authors would like to acknowledge the support of HRV (Changé, France) for their collaborative development of the Virteasy dental simulator software in order to support this project.

Word Count: 3246

MR. JONATHAN DIXON (Orcid ID : 0000-0002-3499-175X)
MR. ASHLEY CHARLES TOWERS (Orcid ID : 0000-0002-2711-1160)
DR. JAMES CLARK FIELD (Orcid ID : 0000-0002-5462-4156)

Article type : Original Article

Re-defining the virtual reality dental simulator: Demonstrating concurrent validity of clinically relevant assessment and feedback

Abstract

Introduction

Virtual reality (VR) dental simulators are gaining momentum as a useful tool to educate dental students. To date, no VR dental simulator exercise has been designed which is capable of reliably providing validated, meaningful clinical feedback to dental students. This study aims to measure the concurrent validity of the assessment, and the provision of qualitative feedback, pertaining to cavity preparations by VR dental simulators.

Methods

A cavity preparation exercise was created on a VR dental simulator, and assessment criteria for cavity preparations were developed. The exercise was performed 10 times in order to demonstrate a range of performances and for each, the simulator feedback was recorded. The exercises were subsequently three-dimensionally printed and 12 clinical teachers were asked to assess the preparations according to the same criteria. Inter-rater reliability (IRR) between clinical teachers was measured using a free-marginal multirater kappa value. Clinical teacher assessment responses were compared with the VR simulator responses and percentage agreements calculated.

Results

IRR values for each exercise ranged from 0.39-0.77 (69.39-88.48%). The assessment of smoothness (κfree0.58, 78.79%) and ability to follow the outline (κfree0.56, 77.88%) demonstrated highest agreement between clinical teachers, whilst the assessment of undercut (κfree0.15, 57.58%) and

depth (κfree 0.28, 64.09%) had the lowest agreement. The modal percentage agreement between clinical teachers and the VR simulator was, on average, 78% across all exercises.

Conclusion

The results of this study demonstrate that it is possible to provide reliable and clinically relevant qualitative feedback via a VR dental simulator. Further research should look to employ this technique across a broader range of exercises that help to develop other complex operative dental skills.

Introduction

Dental students must be capable of carrying out basic operative dental procedures prior to treating real patients safely and effectively [1,2]. Many of these skills are complex to learn, involving the acquisition and application of knowledge and the development of fine motor skills. Pre-clinical operative dental training is commonly carried out within a clinical skills laboratory and within Europe, the vast majority of these are equipped with mechanical patient simulators, commonly referred to as "phantom-heads" [3]. These phantom heads exist typically as replicas of a human head and torso, fitted with jaws that contain either extracted human, or plastic typodont teeth. Phantom heads are used as a basis for both teaching and assessing the necessary operative techniques in order to demonstrate that students are safe to progress to treat patients. Despite the ubiquitous nature of the clinical skills laboratory, the construct is resource-intensive, in terms of time, staffing, restorative materials and tooth substrates [4].

In Dentistry, Virtual Reality (VR) simulators are computer-based systems that attempt to recreate aspects of the real world and often incorporate physical interactivity through haptic technology that provides tactile force-feedback to the user. VR simulators have been successfully employed in the learning of high-risk procedures in aviation and surgery [5,6]. These systems are gaining momentum as a useful tool to educate dental students [7,8]. The reported advantages of VR simulation in dental education include [9-11]:

- the potential to provide iterative and unlimited practical learning
- greatly reduced overheads for resource consumables and teaching staff
- immediate, objective feedback
- the ability to create tailored and standardised exercises

It is clear that VR simulators have the potential to complement traditional teaching methods in pre-clinical operative skills training. However, it is important to recognise that VR simulators need to be supported by well-defined and clear pedagogic values in order to maximise their utility - and this includes validated approaches to assessment [12,13].

The validity of VR systems

Validity can be defined as "*the extent to which an assessment instrument measures what it was designed to measure*" [14]. Different aspects of validity can be demonstrated through objective (construct, concurrent and predictive validity) or subjective (face and content validity) means [15]. Most of the literature that attempts to establish the validity of VR dental simulator feedback claims

to establish *construct* validity, by comparing the assessment of the performance of experts and novices [11,16-18]. Most often this involves comparison of single criterion data, although it is argued that a number of different sources of evidence are required in order to demonstrate and establish construct validity [19-21]. Other studies have attempted to establish the *predictive* validity of their simulator feedback by comparing student performance with a VR simulator and subsequently, after a time lag, with traditional pre-clinical course performance [22,23].

To date, there is no published research that attempts to validate simulator feedback for an exercise by comparing it to an externally validated measure of the same performance. This is known as *concurrent* validity and would involve comparing simulator feedback to that of a trained clinical tutor. A likely reason for this lack of research is that all of the published assessment methods in VR dental simulators are quantitative in nature [4,9,11,16-18,23,24]. The exercises that have been developed for dental education typically involve either the preparation of various geometric shapes [16,18,23,24], or operative procedures on teeth [4,9,11,17,22,25,26]. This quantitative feedback typically provides the user with a score that is based on the amount of the target material removed, the amount of surrounding (non-target) material removed and the time taken to complete the exercise.

Quantitative feedback is often considered advantageous [4,9,18,26,27], primarily due to the objectivity that it provides. However, the true usefulness of this quantitative feedback is questionable as the scoring model is not truly reflective of the task or domain structure itself. For example, the presentation of a coloured region of tissue to be removed provides a clear indication of what is expected within the exercise – although the score does not reveal if a good performance is as a result of a sound understanding of the principles of cavity design, or simply the operator having a steady hand. This is known as construct-irrelevant easiness [21]. In clinical settings, students receive *qualitative* feedback on their performance, which should be meaningful and actionable to support students in improving their performance. Examples of such feedback for an occlusal cavity may include: handpiece control, depth of the preparation and flatness of the floor of the preparation [28].

Despite multiple calls for feedback to conform with that given by dental educators in clinical settings [11,18], to date no VR dental simulator exercise has been designed which is capable of reliably providing this meaningful clinical feedback. This sentiment is echoed by Bakr [29] and Rhienmora [4]. In reality, designing VR software that provides qualitative clinically relevant feedback is undoubtedly extremely complicated, and this may be the primary reason for its underdevelopment.

Aims

This study aims to:

- introduce a novel process for measuring aspects of the validity of the assessment provided by VR dental simulators
- demonstrate a proof of concept for the provision of qualitative clinical feedback with VR dental simulators
- demonstrate the concurrent validity of the VR dental simulator feedback by comparing it with that obtained from clinical tutors

Methods

A visual outline of the methods is presented in Figure 1. An exercise that focussed on the essential features of occlusal cavity preparations was conceptualised by the authors, and developed for use on a Virteasy dental simulator by HRV (Changé, France). The exercise consisted of a block of material having a simulated density similar to human enamel and had a straight-line template on its surface. Users were asked to prepare a cavity of 2mm depth, with maximum undercut, that followed the line. The instruments available for the exercise consisted of a high-speed dental handpiece, a pear-shaped diamond bur and a dental probe. A screenshot of the exercise can be seen in Figure 2.

Assessment criteria and feedback statements

Objective and qualitative criteria for assessing the preparation were obtained from existing published teaching material [28,30]. These criteria were combined with a range of feedback statements derived from published teaching material [28,30] and the expert opinions of experienced senior clinical teaching staff within the School of Clinical Dentistry, University of Sheffield, UK. The qualitative assessment criteria and associated feedback statements can be seen in Tables 1 and 2, respectively.

Development and testing of the assessment and feedback

Software modifications were made to the Virteasy simulator to enable it to make judgements about each of the qualitative assessment criteria based on user performance on the exercise. This involved empirical refinement of mathematical rules and thresholds based on user motions and handpiece angulation until the simulator analysis was aligned with each of the qualitative assessment criteria.

The methods of calculation that the simulator employed for each qualitative assessment criteria are summarised in Table 3. Based on the output of these measurements, threshold values were set to determine a "yes" or "no" judgement for each criteria. This allowed the simulator software to quantitatively assess a preparation, and yet provide qualitative feedback statements to the user.

Once these methods of calculation were established and the exercise was able to provide qualitative statements across the five assessment criteria, a period of testing was undertaken to ensure the simulator always provided the expected feedback. This testing involved the repetitive assessment of preparations of varying quality and a comparison between the clinician's judgement of the preparation and that provided by the simulator. The threshold for each of the methods of calculation were modified until the simulator analysis was aligned with expected clinical feedback, as agreed by the clinical members of the project team (JD, JF, NM).

The delivery of feedback

Once the exercise is completed, users are asked to critically appraise their own work across the 5 cavity features (Table 1, Figure 4). The simulator then delivers its assessment of the actual performance along with any necessary recommendations for improving the performance (advice statements in Table 2) alongside the user's assessment of their own work. This should encourage critical reflection about any discrepancies in the user's perceived performance and the objective assessment of the simulator.

The validation procedure

To establish the concurrent validity of the assessment provided by the simulator, the obtained qualitative statements were compared to clinical teachers' assessment of the same preparations (as the standard). A series of 10 attempts at the exercise were produced by the project clinical skills lead (JD) in order to specifically demonstrate a range of good and bad performances based on the identified assessment criteria presented in Table 1. A combination of preparation errors were prescribed across the 10 exercises (Table 4). For each of these 10 exercise attempts (A-J), the simulator's assessment (yes or no) for each of the 5 assessment criteria was recorded.

Concurrently, the 10 exercise attempts were exported in stereolithography (STL) format and three-dimensionally (3D) printed in the same dimensions using a stereolithography (SLA) 3D printer (Form 2 - Formlabs, Somerville, Massachusetts, USA). A separate overlay template showing the correct

position of the straight line was printed in clear resin to facilitate assessment of the user's ability to follow the outline. An example of the 3D printed models can be seen in Figure 3.

Data collection

In order to assess the 3D printed models, assessors were equipped with a straight probe and magnification as per individual routine practice, plus the transparent position template.

12 clinical teachers were asked to assess each preparation, based on the same criteria as the VR simulator (Table 1). The clinical teacher's assessments were blinded from the VR simulator assessment scores and the project clinical skills lead (JD), who produced the preparations, did not assess the preparations.

Statistical Analysis

The inter-rater reliability (IRR) for assessment scores between the clinical teachers determined by measuring a free-marginal multirater Kappa value, as described by Randolph [31]. This test was chosen to account for the fact that examiner distributions of scores into categories was not restricted. The IRR was calculated per exercise and for each assessed criteria (cavity feature). Exercises that demonstrated low (<0.3) free-marginal multirater kappa scores for IRR were excluded from further agreement analyses with the VR simulator scores.

In order to validate the VR simulator feedback, pooled clinical teacher assessment responses were compared with the VR simulator responses and percentage agreements were calculated. The mode of clinical teacher responses for each assessment criteria for each exercise was also calculated. This allowed for comparison between the "average" clinical teacher and the VR simulator assessments through percentage agreements.

Results

The IRR per exercise, calculated as the free-marginal multirater kappa and the percentage of inter-rater (IR) agreement, can be seen in Table 5. The IRR for two exercises (C,D) fell below the 0.30 κ_{free} score threshold and were subsequently removed from further analyses. The κ_{free} values for the remaining exercises ranged from 0.33-0.77, with the percentage agreement ranging from 66.36-88.48%.

The IRR per assessment criteria (cavity feature), calculated as the free-marginal multirater kappa and the percentage of inter-rater agreement, can be seen in Table 6. The κ_{free} values for the assessment criteria ranged from 0.15-0.58, with the percentage agreement ranging from 57.58-78.79%. The assessment of smoothness of the preparation (κ_{free}0.58 78.79%) and the ability to follow the outline (κ_{free}0.56, 77.88%) demonstrated the highest agreement between clinical teachers. Whilst, the assessment of undercut (κ_{free}0.15, 57.58%) and depth (κ_{free} 0.28, 64.09%) demonstrated the lowest agreement between clinical teachers.

The degree to which the pooled clinical teacher assessments agreed with the VR simulator's assessment was then analysed. This is reported as a percentage agreement with the simulator, per exercise (Table 7). The percentage agreement of clinical teachers and the VR simulator ranged from 40.00%-93.33% depending on the exercise assessed, with a mean score of 70.83% agreement across all exercises. Exercises A and H demonstrated the highest agreement between clinical teachers and the VR simulator, with 93.33% and 85% agreement respectively. Exercises F and I demonstrated the lowest agreement, with 48.33% and 40% respectively.

Given that we expected a degree of variance in the clinical teachers' responses, the *modal* response (agree or disagree) for each assessment criteria and exercise, was then compared to the VR simulator assessment (Table 8). These agreements ranged from 20-100% depending on the exercise. The mean agreement across all exercises was 77.5%. Similar to the pooled data, exercises A (100%) and H (100%) demonstrated high agreement, whilst exercise F (40%) and I (20%) demonstrated the lowest agreement across the two assessors.

Discussion

Currently, there is no published evidence that VR dental simulators are able to provide validated, qualitative feedback in a manner akin to that provided by dental educators in a clinical setting. Whilst there have been attempts to establish the construct validity of VR dental simulators by comparing the performance of expert and novice dental professionals [11,16-18] , it is not clear how useful existing computer-derived quantitative feedback is to students. Repetitive practical experience might result in improvements in the performance of completing a specific task as measured by objective criteria - in the same way that expert dentists might perform better than novices. However, these task-specific percentage scores are more a measure of 'shape agreement',[8] i.e. how well the user can control the handpieces to follow a predetermined pattern. Whilst there may be a degree of demonstrable correlation with this approach[11,16-18,26], this feedback does not

relate or translate to other operative clinical tasks or reflect the structural aspects of the construct domain [21].

Carter [32] argues that meaningful and clinically relevant feedback is a vital part of the learning process. Some examples exist of VR exercises that provide feedback in relation to force application and mirror position [17], however these are difficult to standardise in a VR system, and the value of this feedback to learners is questionable. Instead, the authors would argue the need for more 'human' or 'clinical teacher-style' feedback that more closely matches the feedback given within a real clinical environment. Further, this approach is more robust pedagogically, as it indicates to the user how they might improve and supports self-assessment and reflection, the importance of this in improving clinical competence was demonstrated by de Peralta *et al.* [33].

Other authors have used tutors to contribute to the assessment of criterion measurements of their simulators [11,34], by looking for independent corroborative evaluations of performance. However, this paper presents the first example of establishing a measure of external validity of a simulator's feedback approach using the same criteria as used by the simulator itself. The use of 3D prints of the exercise attempts allowed the assessors to evaluate the performances using the tools and approaches that they would normally use in a clinical setting. This facilitates a more authentic feedback process and mitigates against the confounding factors which might be caused by the differences between the VR environment and the real-world[8].

A high level of agreement was demonstrated between clinical teachers and the simulator after removal of two exercises that had low IRR. As it would not be appropriate to assess simulator agreement with an exercise that a group of experienced dental educators could not agree on, a decision was made by the authors to remove exercises that had low IRR and a threshold was set at <0.3 free-marginal multirater kappa score [35]. The decision to remove these exercises from further analyses was taken to ensure that these analyses were comparing the simulator assessment with clinical teachers that showed a fair to moderate level of agreement. This point brings to light an unexpected level of poor correlation with some tasks; a point that will require further investigation in the development of these validation criteria. After the removal of exercise C and D, the free-marginal multirater kappa scores for the IRR between clinical teachers demonstrated fair to moderate agreement at a minimum [35].

It is important to highlight that clinical teachers who assessed the preparations, did so in the manner of a routine clinical teaching assessment and were not specifically calibrated to assess these exercises. Whilst calibration of assessors may have led to an increased IRR across all of the exercises and cavity features, the authors felt that calibration for a routine operative dental exercise (assessed against standardised features) would not be representative of a routine assessment of operative skills. Furthermore, a degree of variance is expected between clinical teachers even when assessing preparations against objective criteria, and this phenomenon is reported by Seet *et al*[36]. As such, we expected that obtaining high levels of agreement between the clinicians and the VR simulator would be challenging. Despite this, the results demonstrated a mean agreement across all exercises of 77.5%.

Higher than average (over 80%) agreement between the clinical teachers and the simulator was obtained for exercises A, B, H and J. Interestingly, these exercises demonstrated the extremes of each set of criteria; these results are expected, and suggest that clinical teachers and simulators are more likely to agree when a preparation is more obviously "good" or "bad". Exercises that showed the lowest agreement between clinical teachers and the simulator (I and F), demonstrated similar errors with the preparations. These consisted of the preparations being too deep, having insufficient undercut and not being smooth enough. This finding is in agreement with the IRR scores per cavity feature, and, anecdotally with the authors' experience, that depth and undercut appear to be the most challenging of the criteria to reliably assess. The finding is also in keeping with Seet *et al*[36] who reported that less obvious features of crown preparations (such as occlusal reduction) resulted in lower inter-rater agreement than features that were more easily assessed (such as marginal width). Here, the kappa values reported for IRR were significantly lower, ranging from $\kappa = 0.103$ (slight agreement) to $\kappa = 0.399$ (fair agreement). The remaining exercises in this study (E, G) showed strong agreement and incidentally only contained one of the two challenging criteria described above (undercut). Finally, the results suggest that it is the more borderline performances that result in greater disagreement between clinical teachers. This is also expected and demonstrates the true value of the simulator scores in these cases - in order to ensure consistent feedback is delivered to students. It also highlights the importance of the data analysis thresholds that are set for exercise analysis and feedback.

The statistical methods used in this study were carefully chosen to match a relatively complex data set. A free-marginal multirater kappa (κ_{free}) was used to measure the IRR due to the number of assessors; the commonly used Cohen's kappa is only designed for two raters [35]. The κ_{free} was also

selected due to the complexity involved in each assessors assessing all five independent criteria (cavity features) per exercise. When comparing clinical teacher and simulator agreements, the use of percentage agreement is a suitable test - and it is particularly useful when the responses are limited to two values (yes or no) [35].

Whilst the results of this study are very promising in terms of showing that a simulator can generate clinically relevant feedback based on assessment criteria comparable to those used by a tutor, this novel method of assessment and feedback is currently limited to a single simple exercise - as such, further research must look to employ this technique across a broader range of exercises that help to develop other complex operative dental skills. This method of objective qualitative assessment and feedback will be of particular value in relation to feedback criteria that typically generate low tutor IRR.

This proof-of-concept study has demonstrated that clinically relevant, *qualitative feedback* is possible with VR dental simulators. This was achieved by establishing assessment criteria and corresponding qualitative feedback statements for dental operative skills exercises, linking them to measurements on computer systems and subsequently comparing the assessment given by the simulator with dental clinical teachers. The results of this study demonstrated a high level of agreement between clinical teacher assessment and that provided by the VR dental simulator. This suggests that, for the exercise used, it is possible for simulators to reliably assess and provide valid, meaningful and qualitative feedback to students on their performance.

Conclusion

The results of this study demonstrate that it is possible to provide reliable and clinically relevant qualitative feedback via a VR dental simulator. These findings provide a proof of concept for the concurrent validity of VR dental simulator assessment by comparing it to dental educator assessment. Further research should look to employ this technique across a broader range of exercises that help to develop other complex operative dental skills.

Data availability statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

1. Field JC, Cowpe JG, Walmsley AD. The Graduating European Dentist: A New Undergraduate Curriculum Framework. *European Journal of Dental Education.* 2017;21(S1):2-10.
2. General Dental Council. Preparing for practice: Dental team learning outcomes for registration. 2015 revised edition. https://www.gdc-uk.org/docs/default-source/quality-assurance/preparing-for-practice-(revised-2015).p
   Accessed 3rd April 2020.
3. Field J, Stone S, Orsini C, et al. Curriculum content and assessment of pre-clinical dental skills: A survey of undergraduate dental education in Europe. *Eur J Dent Educ.* 2018;22(2):122-127.
4. Rhienmora P, Haddawy P, Suebnukarn S, Dailey MN. Intelligent dental training simulator with objective skill assessment and feedback. *Artif Intell Med.* 2011;52(2):115-121.
5. Allerton DJ. The impact of flight simulation in aerospace. *The Aeronautical Journal (1968).* 2010;114(1162):747-756.
6. Issenberg SB, McGaghie WC, Petrusa ER, Lee Gordon D, Scalese RJ. Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. *Med Teach.* 2005;27(1):10-28.
7. Roy E, Bakr MM, George R. The need for virtual reality simulators in dental education: A review. *The Saudi Dental Journal.* 2017;29(2):41-47.
8. Towers A, Field J, Stokes C, Maddock S, Martin N. A scoping review of the use and application of virtual reality in pre-clinical dental education. *Br Dent J.* 2019;226(5):358-366.
9. Ria S, Cox MJ, Quinn BF, San Diego JP, Bakir A, Woolford MJ. A Scoring System for Assessing Learning Progression of Dental Students' Clinical Skills Using Haptic Virtual Workstations. *J Dent Educ.* 2018;82(3):277-285.
10. Perry S, Burrow MF, Leung WK, Bridges SM. Simulation and curriculum design: a global survey in dental education. *Aust Dent J.* 2017;62(4):453-463.
11. Suebnukarn S, Chaisombat M, Kongpunwijit T, Rhienmora P. Construct validity and expert benchmarking of the haptic virtual reality dental simulator. *J Dent Educ.* 2014;78(10):1442-1450.
12. Flanagan B, Nestel D, Joseph M. Making patient safety the focus: crisis resource management in the undergraduate curriculum. *Med Educ.* 2004;38(1):56-66.
13. Salas E, Bowers CA, Rhodenizer L. It is not how much you have but how you use it: toward a rational use of simulation to support aviation training. *Int J Aviat Psychol.* 1998;8(3):197-208.

14. Van Nortwick SS, Lendvay TS, Jensen AR, Wright AS, Horvath KD, Kim S. Methodologies for establishing validity in surgical simulation studies. *Surgery.* 2010;147(5):622-630.
15. McDougall EM. Validation of surgical simulators. *J Endourol.* 2007;21(3):244-247.
16. Ben-Gal G, Weiss EI, Gafni N, Ziv A. Testing manual dexterity using a virtual reality simulator: reliability and validity. *Eur J Dent Educ.* 2013;17(3):138-142.
17. Suebnukarn S, Haddawy P, Rhienmora P, Jittimanee P, Viratket P. Augmented kinematic feedback from haptic virtual reality for dental skill acquisition. *J Dent Educ.* 2010;74(12):1357-1366.
18. Mirghani I, Mushtaq F, Allsop MJ, et al. Capturing differences in dental training using a virtual reality simulator. *Eur J Dent Educ.* 2018;22(1):67-71.
19. Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin.* 1954;51(2, Pt.2):1-38.
20. Cronbach LJ, Meehl PE. Construct validity in psychological tests. *Psychological Bulletin.* 1955;52(4):281-302.
21. Messick S. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist.* 1995;50(9):741-749.
22. Imber S, Shapira G, Gordon M, Judes H, Metzger Z. A virtual reality dental simulator predicts performance in an operative dentistry manikin course. *Eur J Dent Educ.* 2003;7(4):160-163.
23. Urbankova A, Eber M, Engebretson SP. A complex haptic exercise to predict preclinical operative dentistry performance: a retrospective study. *J Dent Educ.* 2013;77(11):1443-1450.
24. Al-Saud LM, Mushtaq F, Allsop MJ, et al. Feedback and motor skill acquisition using a haptic dental simulator. *Eur J Dent Educ.* 2017;21(4):240-247.
25. Quinn B, Cox M. Assessing Student Products to Determine their Clinical Performance Process Skills: a Mixed Method Approach. In: *AERA Annual Conference 2019.*2019.
26. Ashtari P, Cox MJ, Quinn BFA. The Impact of Innovative Haptic Technologies on Dental Assessment. In: *ADEE/ADEA Shaping the Future of Dental Education, At London.*2017.
27. Buchanan JA. Experience with virtual reality-based technology in teaching restorative dental procedures. *J Dent Educ.* 2004;68(12):1258-1265.
28. Field J. Pre-Clinical Dental Skills at a Glance. *Wiley-Blackwell*. 2015.
29. Bakr MM, Massey W, Alexander H. Academic Evaluation of Simodont® Haptic 3D Virtual Reality Dental Training Simulator. *Paper presented at: Gold Coast Health and Medical Research Conference.* 2012.

30. Banerjee A, Watson TF. Pickard's Guide to Minimally Invasive Operative Dentistry. United Kingdom: *Oxford University Press*. 2015.
31. Randolph J. Free-Marginal Multirater Kappa (multirater κfree): An Alternative to Fleiss Fixed-Marginal Multirater Kappa. *Advances in Data Analysis and Classification.* 2010.
32. Carter FJ, Schijven MP, Aggarwal R, et al. Consensus guidelines for validation of virtual reality surgical simulators. *Surg Endosc.* 2005;19(12):1523-1532.
33. de Peralta TL, Ramaswamy V, Karl E, Van Tubergen E, McLean ME, Fitzgerald M. Caries Removal by First-Year Dental Students: A Multisource Competency Assessment Strategy for Reflective Practice. *Journal of Dental Education.* 2017;81(1):87-95.
34. Ioannou I, Kazmierczak E, Stern L. Comparison of oral surgery task performance in a virtual reality surgical simulator and an animal model using objective measures. *Conf Proc IEEE Eng Med Biol Soc.* 2015;2015:5114-5117.
35. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb).* 2012;22(3):276-282.
36. Seet, RH, Soo, PR, Leong, KJM, Pang, JJH, Lee, FKF, Tan, MY. Crown preparations by undergraduate dental students: A comparison of conventional versus digital assessment via an intra-oral scanner. J Dent Educ. 2020; 1– 11. https://doi.org/10.1002/jdd.12285

Table 1 - Cavity preparation assessment criteria and possible assessor responses

Qualitative Assessment Criteria	**Assessor Responses**
Does the preparation follow the prescribed outline?	Yes: "Your preparation follows the prescribed outline"
	No: "Your preparation does not follow the prescribed outline"
Is the preparation an appropriate depth?	Yes: "Your preparation is an appropriate depth"
	No: "Your preparation is an inappropriate depth"
Does the preparation have enough undercut?	Yes: "Your preparation has enough undercut"
	No: "Your preparation has insufficient undercut"
Is the floor of the preparation relatively flat?	Yes: "The floor of your preparation is flat"
	No: "The floor of your preparation is sloped"
Is the preparation smooth enough?	Yes: "Your preparation is smooth"

Table 2 - Advice linked to each qualitative feedback outcome

Qualitative Feedback	**Advice statements**
Your preparation does not follow the prescribed outline	Maintain a finger rest to have more control of the handpiece
	Position yourself and the exercise block correctly in order to have a clear vision of the block and the handpiece
	Don't revisit the preparation multiple times
Your preparation is an inappropriate depth	Maintain a finger rest to have more control of the handpiece
	Ensure the bur is fully seated in the block and is not entered deeper than this
Your preparation does not have enough undercut	Ensure the bur is fully seated into the block
	Minimise the number of entry and exit points (one at each extreme of the line)
	Ensure that the bur is aligned perpendicular to the surface
The floor of your preparation is sloped	Try to maintain the bur at a constant depth
	Maintain a finger rest to have more control of the handpiece
The preparation is not smooth enough	Try to maintain the bur at a constant depth
	Maintain a finger rest to have more control of the handpiece

Table 3- The methods of calculation for each qualitative assessment criteria

Qualitative Assessment Criteria	**Objective computational methods employed**
Does the preparation follow the prescribed outline?	Starting point accuracy Average error/deviation from line One-off error/deviation from line
Is the preparation an appropriate depth?	Average depth across preparation Single points exceeding depth threshold
Does the preparation have enough undercut?	Bur angle tangent and bi-tangent Number of complete or partial bur withdrawals Depth of preparation below (shallower) depth threshold of the bur
Is the floor of the preparation relatively flat?	Level of inclination of the line of best fit running through depth points of the preparation
Is the preparation smooth enough?	Standard deviation of depth values excluding entry and exit points

Table 4 - A list of the 10 exercise attempts with their prescribed features in relation to the assessment criteria

Exercise Attempt	**Prescribed Features in Relation to Assessment Criteria**
A	Preparation does not follow the prescribed outline, too deep, insufficient undercut, sloped floor and not smooth enough
B	Appropriate across all criteria
C	Preparation does not follow prescribed outline
D	Insufficient undercut, sloped floor and not smooth enough
E	Preparation too shallow, insufficient undercut, sloped floor
F	Preparation too deep, insufficient undercut, not smooth enough
G	Preparation does not follow the prescribed outline, insufficient undercut
H	Appropriate across all criteria
I	Preparation too deep, insufficient undercut, not smooth enough
J	Appropriate across all criteria

Table 5 - IRR (free-marginal multirater kappa - κfree) and the percentage of IR agreement for each exercise, rated by 12 clinical teachers. Exercises falling below the minimum IRR (0.30) are highlighted.

Exercise Attempt	κ_{free}	**95% CI**	**% IR agreement**
A	0.77	0.46, 1.00	88.48
B	0.39	0.14, 0.64	69.39
C	0.26	0.08, 0.44	63.03
D	0.04	-0.03,0.12	52.12
E	0.54	0.23, 0.85	76.97
F	0.54	0.23, 0.85	76.97
G	0.40	0.08, 0.72	70.00
H	0.47	0.2, 0.74	73.64
I	0.37	0.02, 0.72	68.48
J	0.33	-0.03, 0.68	66.36

Table 6 - IRR (free-marginal multirater kappa - κfree) and the percentage of IR agreement for each cavity feature, rated by 12 clinical teachers.

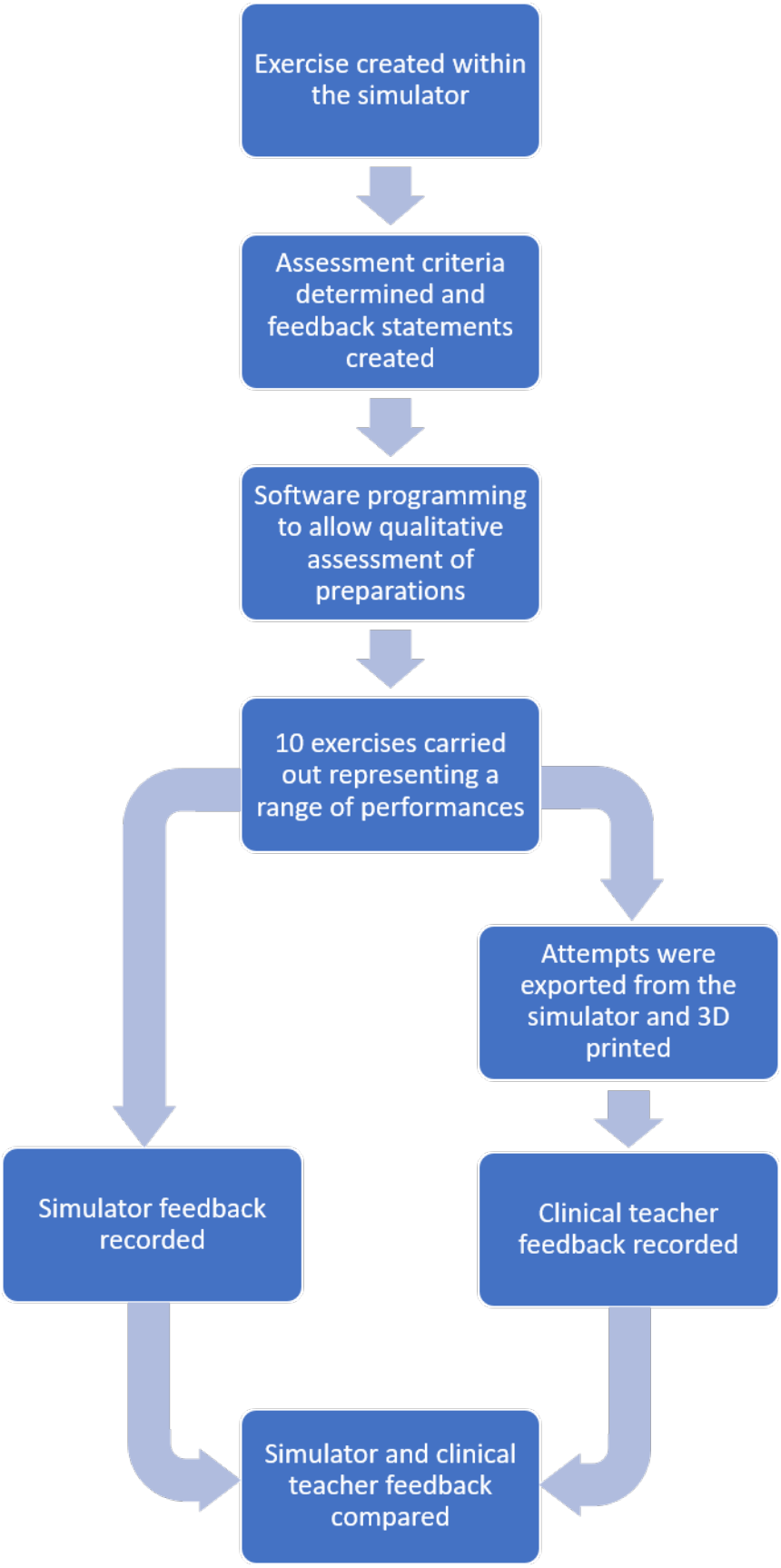
Assessment Criteria	κ_{free}	**95% CI**	**% IR agreement**
Outline	0.56	0.29, 0.83	77.88
Depth	0.28	0.10, 0.46	64.09
Undercut	0.15	0.11, 0.19	57.58
Flat floor	0.46	0.28, 0.65	73.18
Smooth	0.58	0.34, 0.81	78.79

Table 7 - Pooled clinical teacher agreement with simulator score, by exercise

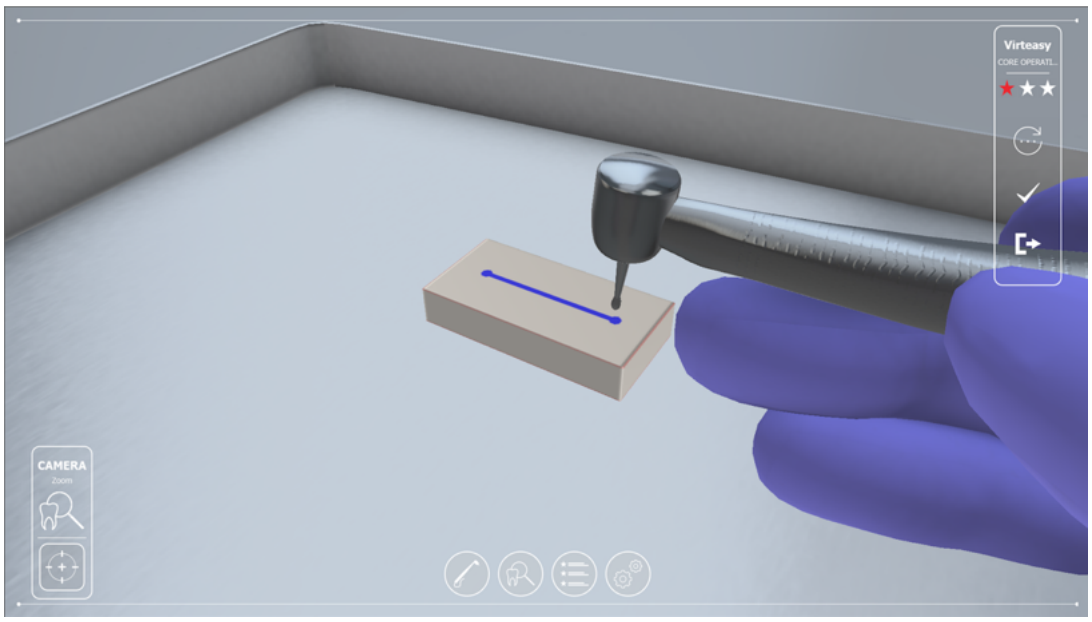
Exercise	Pooled clinical teacher agreements with simulator	Pooled teacher disagreements with simulator	% agreement with simulator
A	56	4	93.33
B	49	11	81.67
E	41	18	68.33
F	29	30	48.33
G	43	17	71.67
H	51	9	85.00
I	24	36	40.00
J	47	13	78.33
Average % agreement with simulator			70.83

Table 8 - Modal clinical teacher agreement with simulator score, by exercise

Exercise	Modal teacher agreements with simulator	Modal teacher disagreements with simulator	% agreement with simulator
A	5	0	100
B	5	0	100
E	4	1	80
F	2	3	40
G	4	1	80
H	5	0	100
I	1	4	20
J	5	0	100
Average % agreement with simulator			77.5

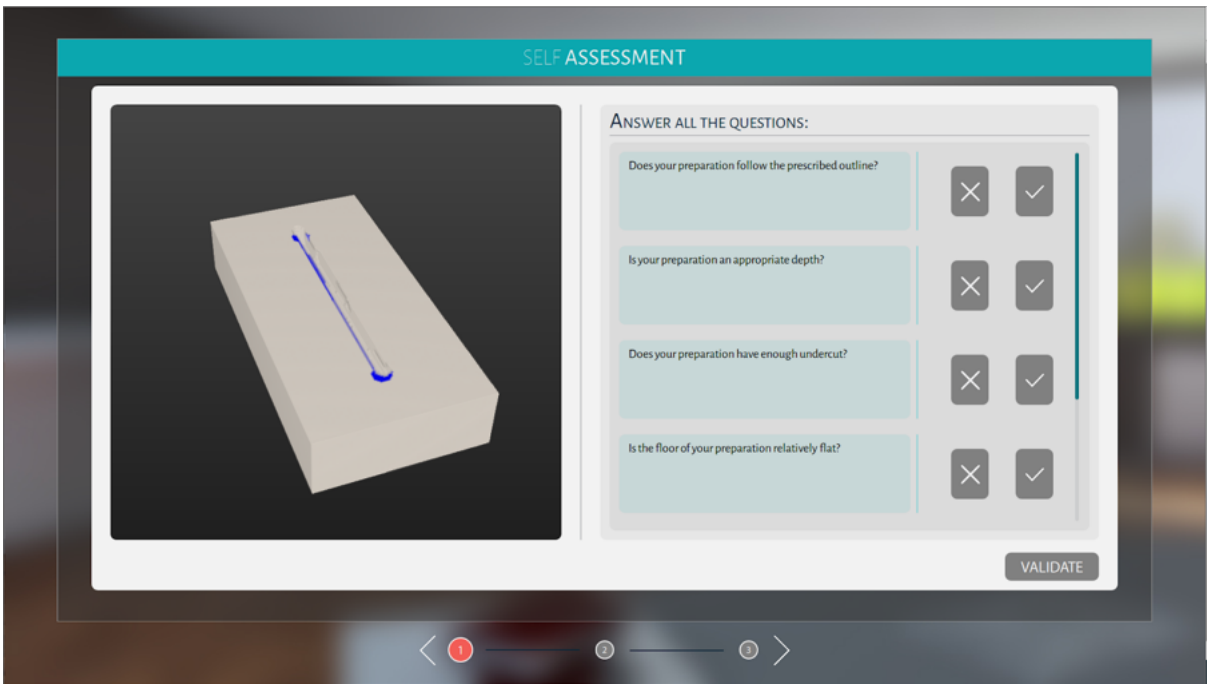


eje_12581_f1.png



eje_12581_f2.png

eje_12581_f3.jpg



eje_12581_f4.png