Received: 21 December 2021 | Revised: 27 October 2022 | Accepted: 31 March 2023

DOI: 10.1111/eje.12908

ORIGINAL ARTICLE

WILEY

Longitudinal assessment of undergraduate dental students: Building evidence for validity

Jamie Dickie[1] | Andrea Sherriff[1] | Michael McEwan[2] | Aileen Bell[1] | Kurt Naudi[1]

[1]University of Glasgow School of Medicine, Dentistry & Nursing, College of Medical, Veterinary & Life Sciences, Glasgow, UK

[2]University of Glasgow, Learning Enhancement and Academic Development Service, Glasgow, UK

Correspondence
Jamie Dickie, University of Glasgow School of Medicine, Dentistry & Nursing, College of Medical, Veterinary & Life Sciences, 378 Sauchiehall St, Glasgow, G2 3JZ, UK.
Email: jamie.dickie@glasgow.ac.uk

Funding information
Dorothy Geddes Research Fellowship; NHS Education for Scotland

Abstract

Purpose: To investigate the content and criterion validity, and reliability of longitudinal clinical assessment of undergraduate dental student clinical competence by determining patterns of clinical performance and comparing them with validated standalone undergraduate examinations.

Methods: Group-based trajectory models tracking students' clinical performance over time were produced from LIFTUPP© data for three dental student cohorts (2017–19; $n=235$) using threshold models based on the Bayesian information criterion. Content validity was investigated using LIFTUPP© performance indicator 4 as the threshold for competence. Criterion validity was investigated using performance indicator 5 to create distinct trajectories of performance before linking and cross-tabulating trajectory group memberships with a 'top 20%' performance in the final Bachelor of Dental Surgery (BDS) examinations. Reliability was calculated using Cronbach's alpha.

Results: Threshold 4 models showed all students followed a single upward trajectory in all three cohorts, showing clear progression in competence over three clinical BDS years. A threshold 5 model produced two distinct trajectories, and in each cohort a 'better performing' trajectory was identified. Students allocated to the 'better performing' trajectories scored higher on average in the final examinations for cohort 2 (29% vs 18% (BDS4); 33% vs. 15% (BDS5)) and cohort 3 (19% vs. 16% (BDS4); 21% vs. 16% (BDS5)). Reliability for the undergraduate examinations was high for all three cohorts (≥0.8815) and did not change appreciably when longitudinal assessment was included.

Conclusions: There is some evidence to support that longitudinal data have a degree of content and criterion validity for assessing the development of clinical competence in undergraduate dental students, which should increase confidence in decisions based on these data. The findings also provide a good foundation for subsequent research.

KEYWORDS
continuous assessment, longitudinal assessment, trajectory, trajectories, undergraduate clinical assessment, validity

1 | INTRODUCTION

A constant challenge faced by educators within healthcare professions is choosing appropriate assessment methods in accordance with the best available evidence to ensure consistent decisions are made in relation to learners' progress. The main purpose of assessment within healthcare professions is to obtain proof that learners have attained the necessary educational outcomes to be certified as safe and competent practitioners.

In the UK, multiple regulatory bodies set professional standards for the training and conduct of healthcare professionals.[1–5] The regulatory bodies maintain registers of individuals who have met these standards and therefore possess the training, skills and experience required to treat members of the public safely and competently.[6] Ultimately, this protects the public from harm.

Dentistry in the UK is regulated by the General Dental Council (GDC), who maintain and publish a list of learning outcomes that must be attained by those seeking entry onto their professional register.[2] These published learning outcomes indicate the necessary knowledge, skills and clinical competencies required to practise dentistry independently in the UK, but they do not stipulate the assessment methods required to demonstrate their attainment. Instead, the responsibility of assessment method selection is delegated to educational institutions.

Assessment methods commonly used within UK dental education include multiple-choice questions (MCQ)/single best answer (SBA) examinations, multiple short answer (MSA)/short answer question (SAQ) examinations, direct observation of procedural skills (DOP)/competency-based tests, objective structured clinical examinations (OSCE), seen and unseen case based examinations[7] and structured clinical reasoning examinations.

Despite their different formats,[8,9] OSCEs, DOPs/competence tests, MCQ/SBA and MSA/SAQs are all types of standalone assessment, which may not be best suited for measuring attainment of some of the GDC's learning outcomes; particularly when evidence of development of *consistent* competent clinical performance is necessary. As a result, some dental schools have incorporated longitudinal clinical assessment into their undergraduate assessment repertoire to allow student knowledge and skills to be evaluated at multiple points in time throughout the curriculum.

Longitudinal clinical assessment has been used within Scottish postgraduate dental vocational training (DVT) schemes for over 10 years. Previous research into its use concluded that it was a 'valid' form of assessment within the context of DVT.[10] However, despite the conclusion of this study and other works which propose that longitudinal data are one of the strongest methods for assessment in clinical subjects,[11–13] their use within undergraduate dental education has yet to be meaningfully investigated.

Assessments that are valid accurately measure what they intend to measure. When investigating validity, the purpose for which an assessment is being used must be defined and the subsequent accumulation of evidence to support its use for this purpose builds a 'validity argument'. Ideally, a validity argument should build evidence on which subtypes of validity can be attributed to the assessment method.[14]

This process could be challenging with respect to longitudinal assessment since the data sets may be very large, with assessment data for multiple skills and attributes measured repeatedly over time for each student. A means of modelling clinical performance over time is required.

In recent years, a form of latent class analysis, known as group-based trajectory modelling (GBTM), has been developed to track groups of individuals following similar patterns of behaviour or achievement of outcome measures over time.[15,16] Although initially developed for use within psychology[17–21] and criminology,[22–24] GBTM is now being increasingly applied to other fields of research, including primary and secondary education,[25,26] physiological medicine[27] and clinical dentistry.[28,29] Therefore, GBTM could be a promising method for analysing longitudinal clinical assessment data produced within dental education.

Briefly, GBTM creates individual student attainment trajectories over a period of time—in this case—by generating lines of best fit to individual student longitudinal clinical assessment data. GBTM then determines whether individual trajectories fit into a smaller number of groups with similar properties. It uses standard statistical methods to choose the optimal number of groups and the shape of the trajectories. As a result, the complexity of the longitudinal data is reduced.

This study primarily seeks to contribute to a validity argument on the use of longitudinal assessment for undergraduate dental student clinical competence. However, it also explores the reliability of various undergraduate assessments used in dental education since—like validity—it is a characteristic of 'good' assessment.[30] The term 'reliability' refers to how reproducible the results of an assessment are.[31] Therefore, highly reliable assessment methods are likely to produce the same or similar results each time they are used.

The aims of this study were to:

1. establish the usefulness of GBTM at modelling longitudinal clinical data and determine patterns of clinical performance over time (content validity— whether the assessment represents what it aims to measure);
2. compare patterns of undergraduate clinical performance from longitudinal assessment with the outcomes of established assessment methods within dental education, namely undergraduate examinations (such OSCE, MSA and MCQ) (criterion validity—how the results obtained via the assessment method correspond to the results of a different test of the same thing); and
3. test the reliability for the panel of assessments used within each BDS year (longitudinal and undergraduate examinations).

2 | METHODS

2.1 | Ethics statement

The study was approved by the University of Glasgow's College of Medical, Veterinary and Life Sciences Ethics Committee for Non-Clinical Research Involving Human Subjects (reference number: 200170146).

2.2 | Design, setting and participants

The graduating classes of the Bachelor of Dental Surgery (BDS) degree programme at the University of Glasgow (UK) from 2017, 2018 and 2019 were eligible for the study. All undergraduate clinical activity during the third (BDS3), fourth (BDS4) and fifth (BDS5) years of the curriculum had been assessed longitudinally for these graduating classes. Longitudinal assessment data were only available from BDS3, BDS4 and BDS5 since these are the years of the course in which University of Glasgow dental students deliver the majority of their patient care.

This resulted in the creation of three longitudinal cohorts: Cohorts 1 (i.e. the class of 2017; $n=80$), 2 (the class of 2018; $n=86$) and 3 (the class of 2019; $n=68$), giving a total sample of 235 participants.

2.3 | Data sources

Assessment data were collected from two sources.

2.3.1 | LIFTUPP©

A digital longitudinal assessment system used to record student clinical activity and performance throughout undergraduate dental training. The system is owned by Turnitin®. All uses of the term 'LIFTUPP©' from this point onwards refer to the assessment system itself.

For each stage of clinical treatment provided, a student is assigned a performance indicator based on a 6-point scale (see Table 1) by supervising clinical faculty (assessor).

Performance indicators and details of the clinical procedure(s) assessed are subsequently uploaded into an extensive database which, by the end of the University of Glasgow's BDS curriculum, typically contains over 100,000 data points of clinical assessment per student cohort and over 1500 data points per student. Each point of clinical assessment records the level of student performance for an individual stage of a clinical procedure.

The three cohorts of LIFTUPP© data were obtained from academic years 2014/15 to 2018/19. These data had previously been collected and stored by the University of Glasgow to monitor student activity and performance with respect to progression and satisfactory completion of the BDS course.

2.3.2 | Undergraduate professional examinations

The results of assessments undertaken by University of Glasgow dental students. The University of Glasgow Dental School currently uses a combination of OSCEs, summative essay assessment, MSA and MCQ examinations to assess the early year groups (i.e. BDS1, BDS2 and BDS3). 'Finals' are assessed with an MSA examination[32] and Comprehensive Care Clinical Case Presentation examination (held in BDS4) and an OSCE[33] (held in BDS5). For this study, the results of these examinations were obtained for the three cohorts between academic terms 2012/13 to 2018/19, and only assessments that were numerically scored were included.

2.4 | Data linkage and pseudonymisation

LIFTUPP© data were stored on an external server by LIFTUPP© Limited (the company owned by Turnitin®). Data management staff at LIFTUPP© Limited transferred data for the University of Glasgow's graduating BDS classes of 2017, 2018 and 2019 to a third-party analyst based at the University of Glasgow via secure email. Undergraduate examination data were transferred to the third-party analyst by the BDS year group administrators using password protection. LIFTUPP© and undergraduate examination data sets were then linked using participant demographics (forename, surname, matriculation number, date of birth and gender) in SAS statistical software (SAS Institute. 2008. SAS software: Release 9.2. Cary, NC: SAS Institute Inc.) by the third-party analyst.

Unique numeric ID codes were then created for each participant before the data were checked for accurate linkage and pseudonymised. The link between the anonymised ID and personal identifiers was stored in a secure drive not accessible to the

TABLE 1 Scale of LIFTUPP© performance indicators and their interpretation.

LIFTUPP© performance indicator	Performance descriptor
1	UNABLE to do this. Has caused harm or does not seek essential guidance.
2	UNABLE to do this independently at present. Largely demonstrated by tutor.
3	UNABLE to do this independently at present but able to complete, to the required quality, with significant help, either procedural or by instruction.
4	ABLE to do this partially independently at the required quality, but requires minor help with aspects of the skill, either procedural or through discussion.
5	ABLE to do this independently at the required quality. This may include confirmatory advice from the tutor where the student seeks appropriate assurance.
6	ABLE to meet the outcome independently, exceeding the required quality.

researchers. Linked (and pseudonymised) data sets were subsequently transferred to a secure drive which was accessible to the researchers.

2.5 | Data processing

Linked data sets were imported into Stata® 15 statistical software (StataCorp. 2017. Stata Statistical Software: Release 15. College Station, TX: StataCorp LLC) and subjected to quality assurance processes. Frequency tabulations and histograms were produced to check ranges and observe any unusual or typographical errors. Cross-tabulations and scatter plots were used to detect logic errors.

Data were cleaned to remove any assisted/observed/simulated activity (leaving only procedures that students had undertaken on patients). Entries with missing data and 'ineligible' procedures were also removed. Procedures were classified as 'ineligible' if an individual student had not performed all stages of a treatment item on an individual patient.

2.5.1 | LIFTUPP© data

Within LIFTUPP©, assessors can assign a performance indicator across four domains: clinical, communication, management and leadership, and professionalism, which are consisting with those outlined by the UK GDC.[2] Data points from the latter three domains were excluded from this study since, between 2014 and 2018, LIFTUPP© did not attribute assessment of these skills to single patient encounters and were assigned per clinical session instead. Some clinical domain data points had also been recorded in this manner—two examples of which were the assessment of extra- and intra-oral examination skills and administration of local anaesthetic. These data points were also excluded.

Therefore, the remaining 'clinical' data points referred to assessment of any hands-on dental procedure recorded as part of a single patient encounter across various disciplines—namely Restorative Dentistry, Paediatric Dentistry, Oral Surgery and Radiology. The list of procedures that were collectively assessed across these disciplines included biopsies, direct restorations, endodontics, extractions, fissure sealants, indirect restorations, minor oral surgery, treatment of temporomandibular disorder (e.g. occlusal splints), periodontal therapy, preventive therapy (e.g. oral hygiene instruction and fluoride applications), removable prosthodontics, radiographs and suturing.

The minimum performance indicator assigned per eligible clinical procedure was used to summarise each student's performance for a clinical procedure at a certain point in time since it was assumed that a student's overall performance would only be as good as their lowest performance indicator for a single procedural stage. Additionally, for each clinical procedure a student undertook, a threshold score of '1' was assigned if the minimum performance indicator was above a threshold level and a 'zero' was assigned if it was below.

Anecdotally, there currently appears to be inconsistency between faculty on whether a performance indicator of 4 or 5 constitutes competent clinical performance. Routine calibration of assessors on the awarding of performance indicators and their relevant criteria had taken place every 6–12 months following the adoption of the LIFTUPP© system at the University of Glasgow Dental School.

For the purposes of this study, performance indicator 4 was used as the baseline threshold for satisfactory clinical performance and to investigate content validity. A threshold performance indicator of 5 was subsequently used to investigate criterion validity, as it allowed more than one trajectory per student cohort to be identified.

2.5.2 | Undergraduate examination data

Average examination performance for individual students was calculated for each BDS year and categorised into fifths. A binary variable was derived for scoring in the top 20% within a year, which aligned with students who had received the best grades (i.e. 'A' grades).

2.6 | Statistical analysis

Summary statistics (mean, standard deviation (S.D), minimum, median, maximum, Q1 and Q3 statistics) were used to describe the number of clinical assessments completed per student per cohort. Frequency tables were produced for the distribution of minimum LIFTUPP© performance indicators awarded per procedure. Mean, S.D, minimum, median, maximum, Q1 and Q3 statistics were also calculated for the minimum LIFTUPP© performance indicators recorded per cohort.

GBTMs based on binary outcomes (i.e. threshold models) were used to model student trajectories for LIFTUPP© data within each cohort year. Models were generated using the traj plugin within Stata statistical software[15] and were based on the probability of participants obtaining threshold performance indicators. The probability of trajectory group membership for each student was calculated and individual students were allocated to the group for which they had the highest probability of membership (i.e. a probability >.5).

For each cohort and threshold performance indicator, a total of 340 GBTMs (in which group trajectories varied by number and/or shape) were generated per LIFTUPP© data set. Models were initially ranked from highest to lowest according to their Bayesian information criterion (BIC), and the five models with BICs closest to zero were analysed for statistical adequacy using statistical tests advocated for GBTMs.[23,34] For models to be statistically adequate, they had to satisfy each of the following criteria:

1. the average posterior probability[35] value is >.70 for each group;
2. the odds of correct classification based on probabilities of group membership is >5 for each group;

3. there is close correspondence between each trajectory group's estimated probability of group membership and the proportion of students classified to that group according to posterior probability of group membership; and
4. tight confidence intervals are observed around estimated group probabilities[23,34]

The BIC of the remaining GBTMs were then compared. According to Raftery's principles, models with a BIC difference greater than two are considered as better fitting and therefore, in accordance with published guidance,[36] were selected as a representative of each cohort's data. If the difference between BICs was less than 2, there is little evidence to support that one model is better than the other.[33] In these instances, the most parsimonious models were selected upon reviewing the graphical representation of each model's trajectory groups.

2.6.1 | Content validity

GBTMs with threshold performance indicator of 4 were used to assess content validity. This is the minimum performance indicator a student can receive to indicate a satisfactory level of clinical performance (see Table 1).

2.6.2 | Criterion validity

For LIFTUPP©, trajectories were compared with already validated undergraduate standalone assessments. To allow these comparisons to be made, more than a single trajectory was needed to distinguish performance between students. A threshold of 4 was not sufficient to produce more than a single trajectory (i.e. all students followed the same trajectory). Therefore, GBTMs using a higher threshold performance indicator of 5 (Table 1) was used to generate more than one trajectory within each cohort. These were then compared with the 'top 20%' of examination performance in each BDS year (1–5) using cross-tabulations with Fisher's exact tests.[37]

2.6.3 | Reliability

Cronbach's alpha[38] was calculated for the panel of assessments between BDS1-5. For LIFTUPP© assessment, the probability of being in the 'best' performing group was used for each student. Cronbach's alpha was then recalculated following removal of each assessment item individually and assessments were considered to be reliable if Cronbach's alpha >0.7.[38]

The sample size for this study was fixed due to only three full cohorts of students' data being available since the commencement of the use of LIFTUPP© at the University of Glasgow Dental School, therefore formal sample size calculations were not possible. Statistical tests were carried out where appropriate, however due to the small numbers, more attention was paid to the effect sizes than the *p*-values.

3 | RESULTS

3.1 | LIFTUPP© data (summary)

Table 2 presents the number of LIFTUPP© assessments eligible for inclusion in the study and the number of students assessed per cohort. It also presents summary statistics for the number of assessments included per student per cohort. Both cohorts 1 and 2 displayed similar means (240.0 and 236.2), minimums (151 and 157) and maximums (349 and 346) per student. Cohort 3 had the largest number of assessments eligible for inclusion collectively (20 817), even though it consisted of fewer students than cohorts 1 and 2. The mean (306.1), minimum (204) and maximum (430) number of assessments were also greater in cohort 3.

Figures 1C display the frequencies of minimum LIFTUPP© performance indicators (per procedure) awarded in each cohort. A score of 5 was the most frequently awarded score in cohort 1 (48%) but (as expected) there was variation across the years. In cohorts 2 and 3, the most common minimum performance indicator awarded to students for clinical procedures was 4 (cohort 2: 46%; cohort 3; 48%). The median minimum LIFTUPP© performance indicator in cohort 1 was 5, whereas in cohorts 2 and 3, it was 4 (Table 3).

3.2 | Content validity

A threshold performance indicator of 4 produced only single group trajectories for all three cohorts (Figures 2–4), which indicates that all students in each cohort follow the same pattern of development over the duration of the BDS course.

TABLE 2 Summary statistics for the number of clinical assessments eligible for inclusion in the study for each cohort and each student within each cohort.

Cohort	LIFTUPP© assessments per cohort (*n*)	Students (*n*)	Mean	SD	Min	Q1	Median	Q3	Max
1	19 199	80	240.0	41.4	151	212.5	240.5	266.5	349
2	20 312	86	236.2	45.8	157	197	232	269	346
3	20 817	68	306.1	47.9	204	270.5	305.5	341	430

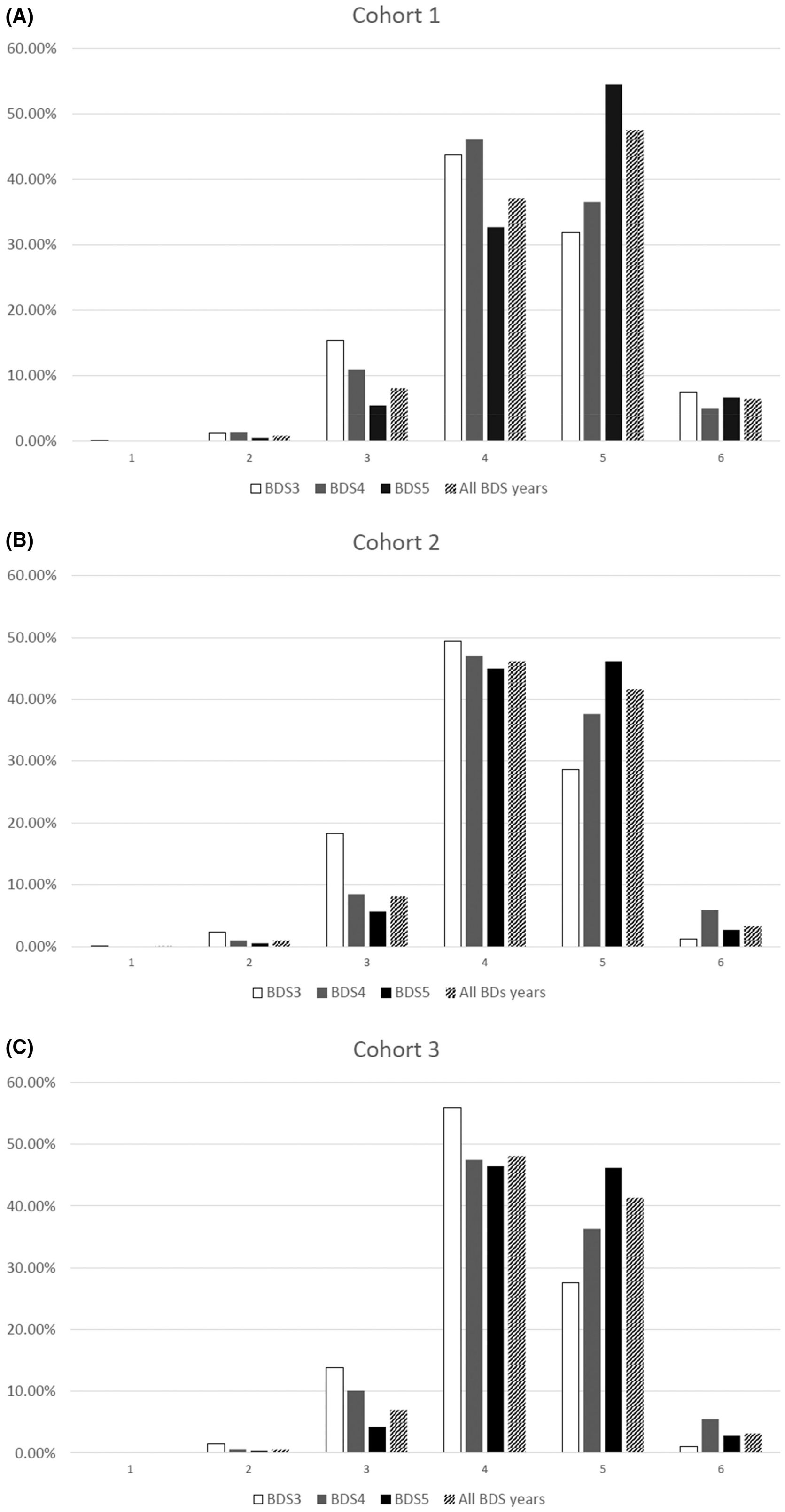

FIGURE 1 Frequencies of minimum LIFTUPP© performance indicators awarded per BDS year and across all BDS years. (A) Cohort 1; (B) Cohort 2; (C) Cohort 3.

TABLE 3 Summary statistics for the (minimum) LIFTUPP© performance indicator recorded for each eligible procedure per cohort.

Cohort	Students (*n*)	Observations (*n*)	Mean score	SD	Min	Q1	Median	Q3	Max
1	80	19 199	4.51	0.77	1	4	5	5	6
2	86	20 312	4.38	0.73	1	4	4	5	6
3	68	20 817	4.40	0.69	1	4	4	5	6

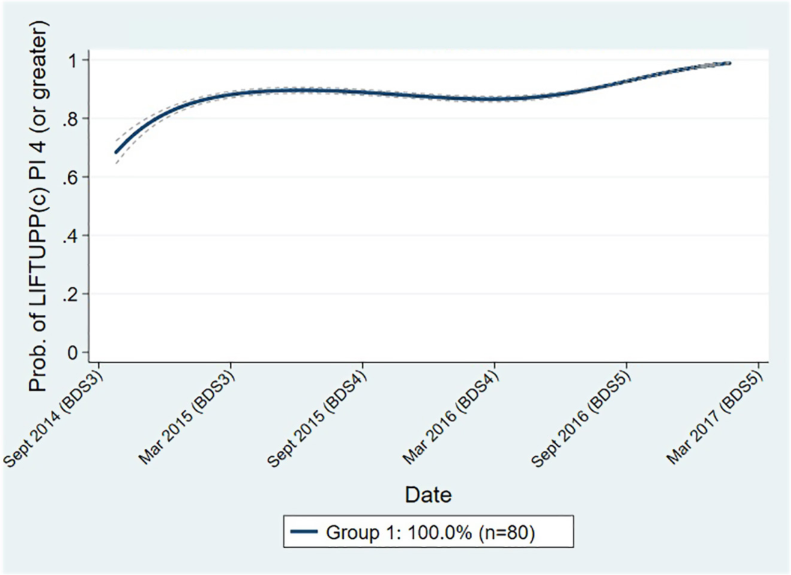

FIGURE 2 Cohort 1 - The trajectory for clinical data if a threshold performance indicator (PI) of 4 was used. The 95% confidence interval around the estimated group probability is depicted by the dotted lines around the trajectory.

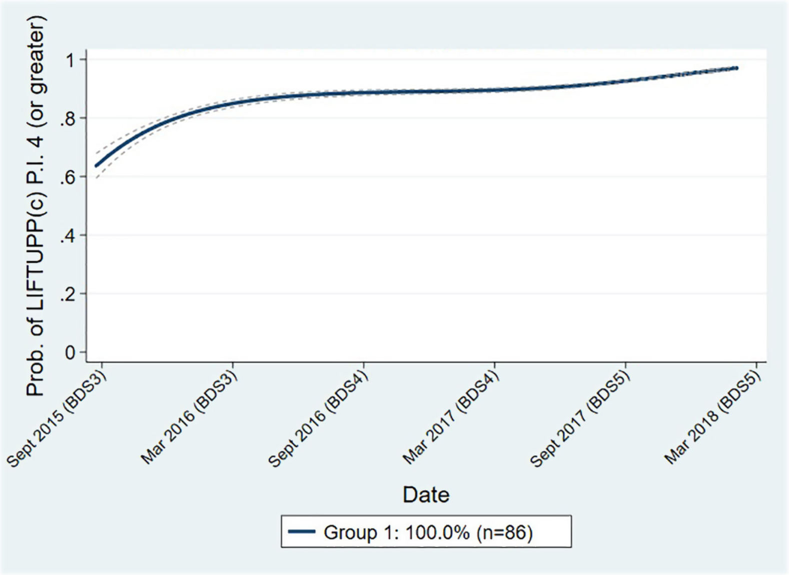

FIGURE 3 Cohort 2 - The trajectory for clinical data if a threshold performance indicator (PI) of 4 was used. The 95% confidence interval around the estimated group probability is depicted by the dotted lines around the trajectory.

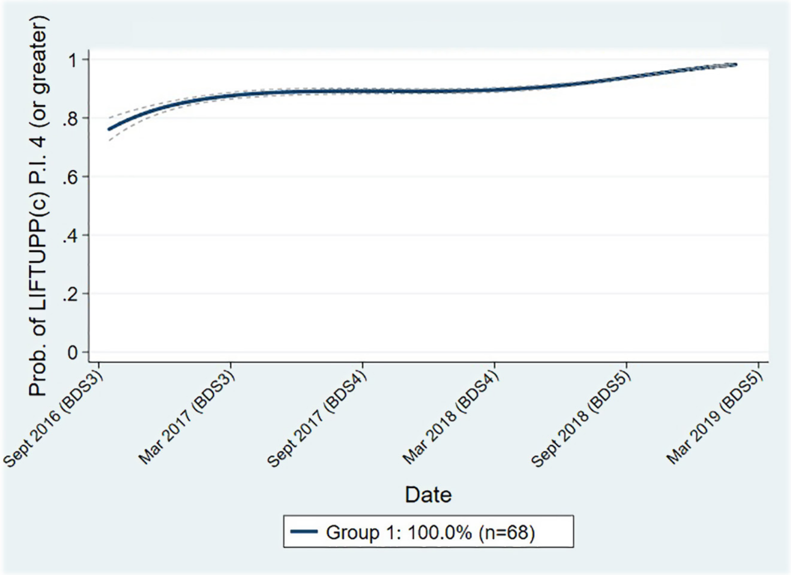

FIGURE 4 Cohort 3 - The trajectory for clinical data if a threshold performance indicator (PI) of 4 was used. The 95% confidence interval around the estimated group probability is depicted by the dotted lines around the trajectory.

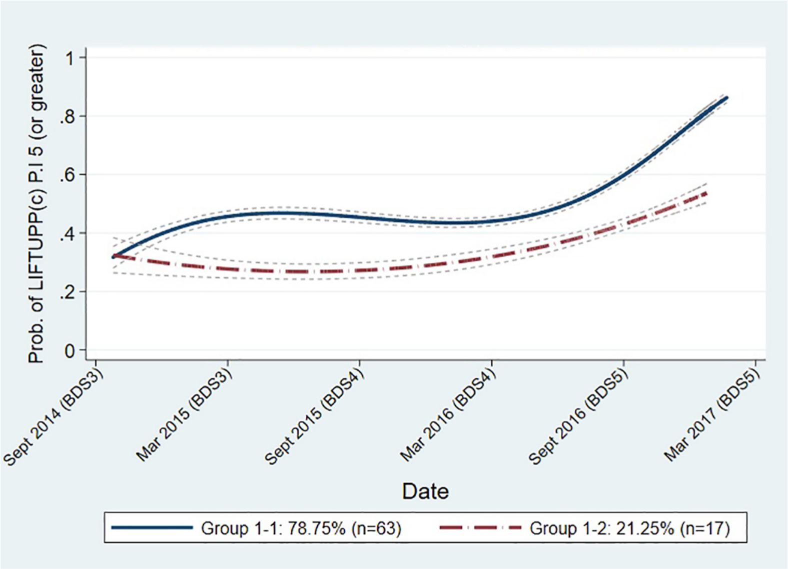

FIGURE 5 Cohort 1 - The trajectories for clinical data if a threshold performance indicator (PI) of 5 was used. The 95% confidence intervals around the estimated group probability are depicted by the dotted lines around each trajectory.

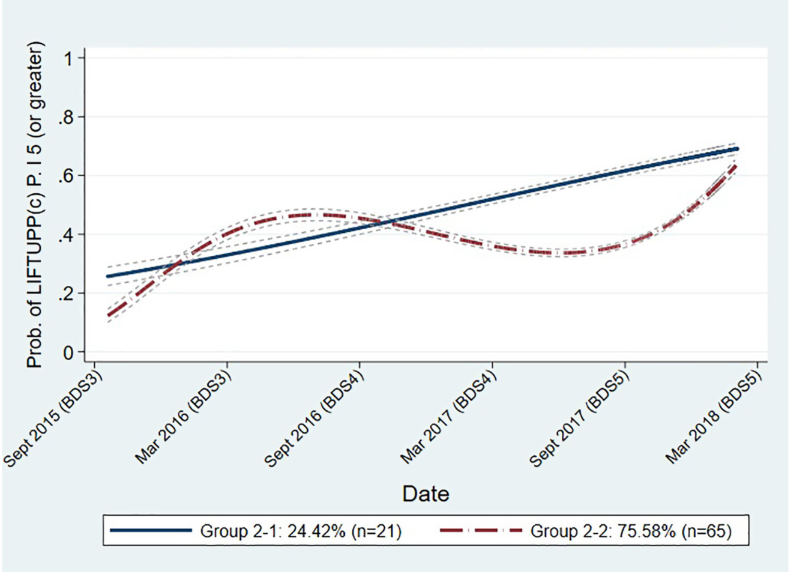

FIGURE 6 Cohort 2 - The trajectories for clinical data if a threshold performance indicator (PI) of 5 was used. The 95% confidence intervals around the estimated group probability are depicted by the dotted lines around the each trajectory. Threshold LIFTUPP© performance indicator 5.

All trajectories showed an increase in probability of achieving LIFTUPP© performance indicator ≥4 between BDS3 and BDS5. By the end of BDS5, the probability of all students in each cohort being assigned ≥4 was over 95%.

Narrow confidence intervals were observed around the estimated group probabilities for the trajectory in each cohort (see dashed lines around trajectories in Figures 2–4).

3.3 | Criterion validity

If a threshold performance indicator of 5 was adopted, threshold models with two distinct trajectory groups fitted the LIFTUPP© data best for all three cohorts (Figures 5–7). Distinction between different groups of student performance meant that models with a threshold performance indicator of 5 could be used for comparisons with other assessment outcomes and, therefore, could be used to investigate criterion validity.

All selected trajectories showed an increase in probability of achieving LIFTUPP© performance indicator ≥5 between BDS3 and BDS5. However, the trajectories differed in starting point, rate of change over time and periodicity.

In cohort 1 ($n = 80$), the first trajectory consisted of 63 students (79%) [group 1–1] and the second of 17 students (21%) [group 1–2] (Figure 5). The first trajectory (solid line) indicates an initial rise in the probability of students achieving LIFTUPP© performance indicator ≥5 between the start and end of BDS3, subsequent plateauing during BDS4, followed by a rapid rise in the probability of achieving performance indicator ≥5 over the duration of BDS5. The second trajectory (dashed line) indicates a longer period of plateau during BDS3 with evidence of a slower rise from BDS4 through BDS5, with no periods of rapid improvement observed. By the end of BDS5, the

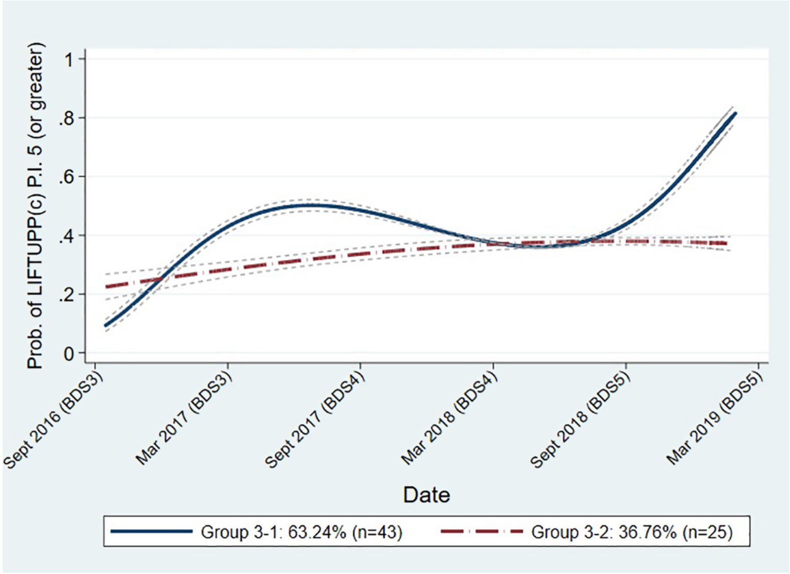

FIGURE 7 Cohort 3 - The trajectories for clinical data if a threshold performance indicator (PI) of 5 was used. The 95% confidence intervals around the estimated group probability are depicted by the dotted lines around the each trajectory.

TABLE 4 Cross-tabulations between LIFTUPP© trajectory group membership and the top 20% of BDS1-5 examination performance.

			Proportion of trajectory group members in top 20% for examination(s) (%)				
Cohort	LIFTUPP© trajectory group	*n*	BDS1	BDS2	BDS3	BDS4	BDS5
1	**1**	63	**23.81**	**23.81**	19.05	17.46	19.05
	2	17	0.00	5.88	**23.53**	**17.65**	**23.53**
	p-value (Fisher's exact)*		.03	.17	.74	1.00	.74
2	**1**	21	**28.57**	**23.81**	**28.57**	**28.57**	**33.33**
	2	65	18.46	20.00	18.46	18.46	15.38
	p-value (Fisher's exact)*		.60	.76	.36	.36	.11
3	**1**	43	**23.08**	**20.93**	**25.58**	**18.60**	**20.93**
	2	25	12.50	20.00	12.00	16.00	16.00
	p-value (Fisher's exact)*		.34	1.00	.23	1.00	.75

Note: Best performing LIFTUPP© trajectory groups and the highest proportion of students in the top 20% for BDS examinations are in bold.

**p*-value generated by Fisher's exact test is based on comparison between proportion of students scoring in top 20% of the undergraduate examiners per BDS year according to LIFTUPP© group membership.

first group had a .85 probability of achieving a performance indicator ≥5 in their LIFTUPP© assessments compared to group 1–2 who had a .55 probability. Overall, group 1–1 appeared to be the better performing group.

The first group of students (solid; 24%; $n=21$) in cohort 2 ($n=86$) [group 2–1] (Figure 6) demonstrated a linear increase in probability of achieving a performance indicator ≥5 from the beginning of BDS3 to the end of BDS5 (.25 to .70). The second group (dashed; 76%; $n=65$) [group 2–2] fluctuated around the linear trajectory over time, dipped slightly during BDS5, but ultimately reached a similar point by the end of the BDS course (with a probability of .65). Group 2–1 appeared to be the better performing group despite both groups almost ending up at the same point by the end of the BDS course.

Finally, in cohort 3 ($n=68$), one group of students (solid; 63%; $n=43$) [group 3–1] (Figure 7) demonstrated an initial increase in their probability of obtaining a LIFTUPP© performance indicator ≥5 over the course of BDS3, a subsequent decrease during BDS4, followed by a rapid rise in BDS5. By the end of BDS5, the probability of group 3–1 obtained a performance indicator of 5 was .81. The second group (dashed; 37%; $n=25$) [group 3–2] also demonstrated an increase in probability but displayed a slower rise across the duration of the BDS course with no periods of rapid improvement. By the end of BDS5, the probability of group 3–2 obtained a performance indicator of 5 was .39. Overall, group 3–1 appeared to be the better performing group.

Narrow confidence intervals were observed around the estimated group probabilities for all trajectory groups in each cohort (see dashed lines around trajectories in Figures 5-7).

The associations between LIFTUPP© group membership based on a performance indicator threshold of 5 and top 20% performance in undergraduate assessments (BDS1-5) in each of the three cohorts are presented in Table 4.

In the two most recent cohorts, a higher percentage of the better performing LIFTUPP© trajectory group consistently produced a top 20% performance in all BDS1-5 undergraduate examinations (Table 4). In cohort 2, the difference in the percentage of students scoring in the top fifth in each year's standalone examinations between the better performing group (group 2–1) and the group who performed less well (group 2–2) were as follows: 28.6% (six of 21) vs. 18.5% (12 of 65) (BDS1); 23.8% (five of 21) vs. 20% (13 of 65) (BDS2); 28.6% (six of 21) vs. 18.5% (12 of 65) (BDS3), 28.6% (six of 21) vs. 18.5% (12 of 65) (BDS4); and 33.3% (seven of 21) vs. 15.4% (10 of 65) (BDS5). For cohort 3, the difference between groups 3–1 (the better performing group) and 3–2 (the group who performed less well group) were as follows: 23.1% (10 of 43) vs. 12% (three of 25) (BDS1); 21% (nine of 43) vs. 20% (five of 25) (BDS2); 25.58% (11 of 43) vs. 12% (three of 25) (BDS3); 18.6% (eight of 43) vs. 16% (four of 25) (BDS4); and 20.9% (nine of 43) vs. 16% (four of 25) (BDS5). This consistent association between the better LIFTUPP© trajectory and top 20% undergraduate examination performance was not observed in cohort 1.

3.4 | Reliability of the assessments

Only LIFTUPP© trajectory models with a threshold performance indicator of 5 could be included as part of the Cronbach's alpha calculations since no discrimination between student clinical performance was found using a threshold performance indicator of 4 (see above).

Overall, the panel of assessments displayed high reliability across all three cohorts since all Cronbach's alpha scores were ≥0.8815 (Table 5). However, inclusion of LIFTUPP© data to the panel of assessments lead to a small decrease in reliability in all three cohorts (Cohort 1–0.9226 to 0.9042; Cohort 2–0.9335 to 0.9246; Cohort 3 0.9152 to 0.8980). Removal of any of the other assessments did not result in a decrease in reliability.

4 | DISCUSSION

This study used a novel means of modelling longitudinal clinical assessment data, and record-linked this data to standalone undergraduate assessment outcomes to demonstrate content and criterion validity for three cohorts of dental students within a single dental school. To the knowledge of the authors, the methodological approach used in this study is the first-time longitudinal data sourced from LIFTUPP© have been modelled in this manner. It is also the first-time LIFTUPP© data have been linked to other forms of assessment used in undergraduate dental education.

Strong data security protocols ensured anonymity and risks to the participants were minimal and there was no impact on their academic records or progression through the BDS curriculum. The data linkage process permitted secondary use of routinely collected assessment data, which is not often utilised within educational research, allowing the relationship between current and emerging assessment practices to be investigated.

Due to the rigorous inclusion criteria (see methods), the number of clinical LIFTUPP© assessments eligible for the study does not reflect the full extent of student clinical experience in each cohort. Procedures were only considered eligible if an individual student performed (and had been assessed on) all stages of a treatment item on an individual patient. Students are timetabled to

TABLE 5 Cronbach's alpha coefficients across all BDS1-5 examinations and LIFTUPP© trajectory group membership probability per cohort.

		Cohort 1		Cohort 2		Cohort 3	
BDS year	Assessment	Observations (*n*)	Cronbach's alpha	Observations (*n*)	Cronbach's alpha	Observations (*n*)	Cronbach's alpha
1	MCQ	90	0.8925	92	0.9152	73	0.8845
	OSCE	89	0.8911	92	0.9187	73	0.8823
2	MCQ	88	0.8942	92	0.9128	72	0.8881
	MSA	88	0.8895	92	0.9140	72	0.8815
	OSCE	88	0.8995	92	0.9245	72	0.8931
3	MSA (anatomy)	87	0.8921	92	0.9138	72	0.8829
	MCQ	87	0.8925	92	0.9167	72	0.8843
	MSA	87	0.8912	92	0.9113	72	0.8824
	OSCE	87	0.8997	92	0.9183	72	0.8919
4	MSA	84	0.8912	91	0.9147	72	0.8950
5	OSCE	82	0.8953	93	0.9236	69	0.8909
3/4/5	LIFTUPP© group prob.	80	0.9226	86	0.9335	68	0.9152
		Overall Cronbach's alpha	0.9042	Overall Cronbach's alpha	0.9246	Overall Cronbach's alpha	0.8980

rotate between different treatment centres (especially during the BDS5 outreach programme) and, as a result, some students will not complete all the relevant stages for procedures spread across multiple visits (e.g. root canal treatments and dentures) on the same patient. One student may complete some of the stages before they are replaced by another student timetabled to attend the treatment centre who completes the remaining stages. Although LIFTUPP© can easily record student experience in performing individual procedural stages (or the components of a procedure), assessment of clinical competence becomes more complicated if assessors are required to piece together parts of different assessments (or procedural stages performed across multiple patients) to determine their true capability in performing clinical tasks. A clearer indication of student clinical competence can be drawn from procedures which have been completed from beginning to end on the same patient, hence why this study opted to focus on procedures completed in their entirety by the same student. Imposing these strict inclusion criteria also facilitated investigation of student competence under the premise that their overall procedural performance would only be as good as their lowest performance indicator for a single treatment stage (see methods section). Additionally, some clinical procedures for cohort 1 were not assessed and recorded during the early stages of LIFTUPP©'s implementation, and, in both cohort 1 and 2, some were not attributable to individual patient encounters and had instead been assigned per clinical session. Procedures affected by these circumstances were discounted via the inclusion criteria and, therefore, the number of procedures included in the study only represent a proportion of the students' clinical experience.

Initial statistical descriptions of LIFTUPP© data demonstrated a clear difference in the total number of clinical assessments eligible for inclusion in the study in cohort 1 compared to cohorts 2 and 3. Although each subsequent cohort had more clinical data that were eligible for inclusion than their predecessors, there was a greater increase between cohorts 1 and 2. Interestingly, cohort 3 completed the most assessments despite consisting of less students. Undergraduate dental year groups at the University of Glasgow typically consist of 65 to 90 students, therefore, despite having a smaller number of students, cohort 3 was not an atypical class size. The increased number of assessments from cohort 3 might be explained by assessors become more accustomed to the LIFTUPP© system over time and/or by drives from educational leads who encouraged assessors to complete as many LIFTUPP© student assessments as possible.

There were also differences in the distribution of LIFTUPP© performance indicators between cohort 1 and cohorts 2 and 3—with the two most recent cohorts recording slightly lower performance indicators. These observations could again be indicative of assessors becoming more familiar with the assessment options available within LIFTUPP© and assigning performance indicators more in line with the guidance. Cohort 1 was the first group of students to have had all undergraduate clinical activity assessed via LIFTUPP© and a 'settling in' period was to be expected. Furthermore, assessor calibration exercises were introduced following the system's adoption. These exercises were conducted every 6–12 months and could have influenced assignment of performance indicators to be more aligned with the guidance for the more recent cohorts since examiner training has been shown to impact scoring.[39] However, even though a difference in scoring patterns was detected, there is currently insufficient strong evidence to support this supposition since the effects of examiner training on LIFTUPP© data patterns was not investigated as part of this study.

Although faculty participate in team calibration exercises in relation to the application of LIFTUPP© descriptors, there is some disagreement between groups of faculty as to which descriptor marks the threshold for competent clinical performance. Some assessors still consider a performance indicator of 4 as the threshold whilst others argue that the threshold is 5.

Assessors were continually encouraged to refer to the LIFTUPP© performance indicator criteria (Table 1) if they were having difficulty distinguishing the difference between performance indicators 4 and 5. They were also shown and guided through examples of student assessment scenarios that highlighted the differences between the awarding of each of the LIFTUPP© performance indicators. However, despite regular calibration of assessors, potential discrepancies in the application of LIFTUPP© performance indicator criteria between assessors could not be completely disregarded. Therefore, in this study, both indicators 4 and 5 have been used to investigate validity. Content validity was investigated using LIFTUPP© performance indicator 4 as the threshold for competence and criterion validity was investigated using performance indicator 5 as this threshold was required to discriminate between the students' performance.

Although there was no distinction in clinical performance between groups of students using performance indicator 4 as the threshold, the upward trend in trajectories between BDS3 and BDS5 suggested students were improving over time. By setting a threshold of 5, it was possible to separate students into two performance groups in each cohort. All the distinct trajectories also demonstrated progression in clinical assessment from BDS3 to BDS5; however, their starting point and rate of change over time differed, with a clearly 'better performing' trajectory within each cohort.

Collectively, the patterns of the trajectories appeared to suggest that LIFTUPP©, and therefore longitudinal data, have a degree of content validity for measuring development of dental student clinical competence since the trajectories reflected patterns of clinical development that would be expected of students as they progress through the curriculum (i.e. an upward trend) and it was possible to detect differing patterns of student performance. However, the lack of distinction between student groups in the models with threshold 4 meant only the threshold 5 models could be used to investigate the criterion validity of LIFTUPP©/longitudinal data.

There was some evidence to suggest that LIFTUPP©/longitudinal data have a degree of criterion validity (in measuring development of clinical performance) since students allocated to the better performing trajectories in cohorts 2 and 3 tended to perform better in standalone undergraduate assessments. There are no 'gold standard' assessment methods within dental education, but the validity of OSCEs[40–46] and MSAs[47] has been well publicised. In addition, although there is less robust evidence to support the validity of MCQs, they remain one of the most frequently used assessment methods within dental education.[7,8,11] OSCEs, MSAs and MCQs form a range of assessments that, together, provides an overall picture of student development and attainment. Therefore, comparing LIFTUPP© data with MCQ, MSA and OSCE outcomes provides a starting point for investigating the criterion validity of longitudinal assessment.

It was worth noting that the low number of LIFTUPP© performance indicators <4 shown in the results may be indicative of a long-standing problem withing dental education—'failure to fail'.[48] This issue occurs when assessors are reluctant to fail students once they are admitted onto the course[49,50] and can occur for a variety of reasons.[51–54]

Even though the LIFTUPP© system seeks to make assessment more objective through provision of descriptors which define the awarding of each performance indicator, there is (once again) no guarantee assessors will always abide by the criteria, or they may interpret the assessment criteria differently based on their own clinical and assessment experience.[55] Berendonk, Stalmeijer and Schuwirth (2013)[56] also suggested assessors may be tempted to make comparisons between individual student performances (i.e. students are benchmarked against one another instead of the assessment criteria). Furthermore, Gingerich, Regehr and Eva (2011)[57] have suggested assessor's judgements can be influenced by a variety of cognitive factors (such as mood, impression formation, and interactions with previous individuals), which again may result in assessors (unconsciously) deviating from the application of strict assessment criteria.

Although investigations into potential occurrences of 'failure to fail' was not an intention of this study, it is important, in the absence of further information, not to rule out this long running issue within dental education as a potential explanation for the returned results as it can lead to serious implications regarding patient safety.[58,59]

Longitudinal assessment formats may have a positive impact on 'failure to fail' if they ensure students are assessed by a range of assessors. This may counter-balance lenient assessors ('doves') against those who are very strict ('hawks')[60–62] and facilitate triangulation of information on performance so an accurate representation of a student's skills and abilities can be formed,[61,63] thus increasing assessment validity.[64–66] However, at present, both the degree of calibration among assessors and the incidence of assessor extremes ('hawks' or 'doves') in relation to longitudinal clinical assessment have yet to be meaningfully explored.

It was noted that the relationship between 'good' LIFTUPP© performance and 'good' examination performance was more pronounced in cohort 2 than in cohort 3 (Table 4). Potential causes for this observation are not yet fully understood but may indicate that trajectory shapes could be influential and require further research. The lack of association in cohort 1 may (again) have been due to varying degrees of calibration among assessors following the initial adoption of LIFTUPP©.

LIFTUPP© appears to be a reliable assessment format. However, incorporating each student's longitudinal clinical performance into a single metric to allow the calculation of a reliability statistic was challenging. Simply assigning a binary value of '1' or '0' to each student did not provide the necessary variation—therefore, the probability of membership of the trajectory was used for each student. This requires further work.

The findings are based on data collected from a single dental school and, due to the relatively recent introduction of LIFTUPP© to the University of Glasgow Dental School, only three cohorts of student data could be analysed. Therefore, as anticipated, the study lacked the power to demonstrate important differences statistically. Despite the small number of participants and cohorts, the LIFTUPP© system provided very large data sets derived from robust (almost real time) electronic data capture. It was possible to apply statistical modelling methods to these data and it appears that GBTM is one which shows promise in modelling the large amounts of time varying data gathered by LIFTUPP© since it provided simple—but meaningful—graphical and tabular summaries. These summaries not only allowed content validity for longitudinal clinical assessment data to be explored, but also facilitated comparisons with other assessments (undergraduate examinations and LEPs) to investigate criterion validity since student trajectory group memberships supplied outcomes of clinical performance.

Like any statistical modelling technique, GBTM has associated limitations, most of which are due to GBTM summarising the average behavioural trend(s) of multiple individuals to create distinct trajectory groups into which each individual can be classified. Data classification processes are susceptible to error[67] and even though GBTM attempts to mitigate errors by basing its classifications on the available data, not every individual allocated to a group will precisely follow the group trajectory. As a result, each trajectory group mainly consists of individuals whose developmental patterns resemble one another (and the overall group trajectory) compared with other trajectories.[24]

There may be a small number of individuals whose data do not follow the more typical trajectories produced by collections of other individuals within the cohort. If there are few individuals following more atypical trajectories, it may not be possible to generate a unique trajectory group for them to be allocated to. Instead, outlying individuals may be approximated into one of the more typical trajectory groups, resulting in some individuals being 'miscategorised'. Therefore, in this study, some students may have been 'upgraded' into a better performing trajectory group or 'downgraded' into a group performing less well, meaning a more accurate representation of their development was lost, which, ultimately, will have influenced data interpretation.

Furthermore, the number, shape and accuracy of the trajectories identified through GBTM are strongly influenced by several factors, such as sample size, number of available data points and the timeframe over which data have been gathered.[24,68] Therefore, the trajectories are not definitively fixed nor are individual memberships to the trajectory groups.[24] Although study power and precision of results can be increased through larger sample sizes,[48] GBTMs will always be approximations of more complex development behaviour(s).[24]

Selection of the two group models as the best fitting (for threshold 5) was predominantly guided by the BIC. In general, models with the BIC closest to zero are the best fitting[19,28] and this also appeared to be the case in this study. None of the models displaying the BIC closest to zero in each cohort were rejected following additional testing for statistical adequacy or because they appeared less parsimonious than other models when their graphical appearance and relationship to the undergraduate results were scrutinised. However, it was possible that more appropriate models may have been missed despite this approach being consistent with guidelines on model selection (as suggested by the developer of the GBTM technique).[23,34]

It is worth noting that, due to the complexity of the analyses and limitations of GBTM, the authors would recommend seeking appropriate statistical support before adopting this modelling technique for further studies.

Since there is currently very little literature on the utility of LIFTUPP© data for assessment of clinical development, it is difficult to relate the findings of this study to others. However, the results presented here echo two previous studies by Prescott-Clements et al.[10] and Roudsari.[7] The former demonstrated upwards trajectories of longitudinal performance over the duration of the UK's 1-year dental postgraduate vocational training period using mean scores recorded by longitudinal evaluations of performance (LEPs)[69] for two cohort of vocational dental practitioners (VDPs). The latter used LIFTUPP© data and an alternative data modelling technique (Bayes theorem) to demonstrate an upward trend in student clinical performance (in oral surgery) over time.[7] Although both the oral surgery study and the study presented in this paper have illustrated trends within LIFTUPP© data, there may be some debate over what these findings ultimately mean. There are several questions which need some consideration, and these include: Does it really matter where students start in their development? Are the ways in which students progress to the end point important? Can these findings be used to enhance assessment?

The opinions of key stakeholders may provide the answers to such questions and contribute further evidence on the validity of longitudinal assessment. However, there is still a need to build the evidence base further. This could partly be achieved through expanding this study across multiple centres and subsequent student cohorts, but the use of other sources and approaches also needs some consideration.

Overall, this study has served as a starting point upon which future studies on the validity of longitudinal data for assessment of the development of clinical skills could be built, not just within dental education, but for other specialties which use, or are considering using, this form of assessment. Providing there are enough observations, the statistical modelling technique used in this study could be applied to individual subject areas within different specialities (e.g., oral surgery within dentistry). This could allow assessors to monitor how students progress over time and identify points at which intervention is required should dips in (average) performance be detected. The minimum number of observations required to ensure accurate longitudinal data trajectory models are generated requires further investigation.

5 | CONCLUSION

There is good evidence to suggest that longitudinal clinical assessment data (recorded via the LIFTUPP© system) have both content and criterion validity for determining development of clinical competence.

With respect to the first objective set out in the introduction section, evidence for content validity was demonstrated by trajectory patterns of clinical performance trending upwards as students progressed through the BDS course, indicating that students' clinical abilities were improving over time.

In terms of the second objective, evidence for criterion validity for longitudinal clinical assessment was illustrated by associations between better undergraduate longitudinal clinical performance and better undergraduate examination outcomes in the two most recent student cohorts (i.e. the graduating classes of 2018 and 2019).

Finally, in relation to the third objective, no conclusive evidence on the degree of reliability for LIFTUPP© data was obtained.

Further studies in different settings (e.g. medicine, nursing, veterinary medicine and other dental schools) and in additional cohorts, as routine use of LIFTUPP© becomes more common, are needed to confirm, and build upon these findings.

ACKNOWLEDGEMENTS

The authors wish to thank Dr Alex McMahon (University of Glasgow) for his assistance and expertise with linkage and pseudonymisation of the data sets used to conduct this study, and Professor Vince Bissell (University of Liverpool) for the conception and initiation of the study. The lead author wishes to thank NHS Education for Scotland and the University of Glasgow for their data, financial and pastoral support and gratefully acknowledges the support of the Dorthey Geddes Scholarship fund.

FUNDING INFORMATION

The lead author received support from the Dorothy Geddes Research Fellowship and NHS Educationfor Scotland to conduct this study as part of a PhD project.

CONFLICT OF INTEREST STATEMENT

The authors of this manuscript have no conflicts of interest to declare.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

ORCID

Jamie Dickie https://orcid.org/0000-0001-8950-3226

REFERENCES

1. General Chiropractic Council. *Education Standards*. General Chiropractic Council; 2017.
2. General Dental Council. *Preparing for Practice (2015 Revised Edition)*. General Dental Council; 2015.
3. General Medical Council. *Outcomes for Graduates*. General Medical Council; 2018.
4. General Optical Council. *Core Competencies for Optometrists*. General Optical Council; 2016.
5. General Osteopathic Council. *Guidance for Osteopathic Pre-Registration Education*. General Osteopathic Council; 2015.
6. UK Health and Safety Executive. *Who regulates health and social care?* [Online]. UK Health and Safety Executive. Accessed July 2020.
7. Roudsari RV. Assessment of competence in dentistry: the expectations, perceptions, and predictions. *Doctor of Philosophy in Clinical Dentistry*. University of Manchester; 2017.
8. Williams J, Baillie S, Rhind S, Warman S, Sandy J, Ireland AA. *Guide to Assessment in Dental Education*. University of Bristol; 2015. Accessed: September 2020.
9. Field JC, Walmsley AD, Paganelli C, et al. The graduating European dentist: contemporaneous methods of teaching, learning and assessment in dental undergraduate education. *Eur J Dent Educ*. 2017;21:28-35.
10. Prescott-Clements L, Van der Vleuten CP, Schuwirth LW, Hurst Y, Rennie JS. Evidence for validity within workplace assessment: the longitudinal evaluation of performance (LEP). *Med Educ*. 2008;42:488-495.
11. Albino JEN, Young SK, Neumann LM, et al. Assessing dental Students' competence: best practice recommendations in the performance assessment literature and investigation of current practices in predoctoral dental education. *J Dent Educ*. 2008;72:1405-1435.
12. Dawson LJ, Mason BG, Bissell V, Youngson C. Calling for a re-evaluation of the data required to credibly demonstrate a dental student is safe and ready to practice. *Eur J Dent Educ*. 2017;21:130-135.
13. Patel US, Tonni I, Gadbury-Amyot C, Van der Vleuten CPM, Escudier M. Assessment in a global context: an international perspective on dental education. *Eur J Dent Educ*. 2018;22(Suppl 1):21-27.
14. Kane MT. Validating the interpretations and uses of test scores. *J Educ Meas*. 2013;50:1-73.
15. Jones BL, Nagin DS. *A Stata Plugin for Estimating Group-Based Trajectory Models*. Carnegie Mellon University; 2012. https://ssrc.indiana.edu/doc/wimdocs/2013-03-29_nagin_trajectory_stata-plugin-info.pdf. Accessed: September 2020.
16. Jones BL, Nagin DS. A note on a Stata plugin for estimating group-based trajectory models. *Sociol Methods Res*. 2013;42(4):608-613.
17. Orcutt HK, Erickson DJ, Wolfe J. The course of PTSD symptoms among gulf war veterans: a growth mixture modeling approach. *J Trauma Stress*. 2004;17:195-202.
18. Dekker MC, Ferdinand RF, Van Lang NDJ, Bongers IL, Van der Ende JD, Verhulst FC. Developmental trajectories of depressive symptoms from early childhood to late adolescence: gender differences and adult outcome. *J Child Psychol Psychiatry*. 2007;48:657-666.
19. Jester JM, Nigg JT, Buu A, et al. Trajectories of childhood aggression and inattention/hyperactivity: differential effects on substance abuse in adolescence. *J Am Acad Child Adolesc Psychiatry*. 2008;47:1158-1165.
20. Mora PA, Bennett IM, Elo IT, Mathew L, Coyne JC, Culhane JF. Distinct trajectories of perinatal depressive symptomatology: evidence from growth mixture modeling. *Am J Epidemiol*. 2009;169:24-32.
21. Girard LC, Tremblay RE, Nagin D, Côté SM. Development of aggression subtypes from childhood to adolescence: a group-based multi-trajectory modelling perspective. *J Abnorm Child Psychol*. 2019;47(5):825-838.
22. Weisburd D, Bushway S, Lum C, Yang SM. Trajectories of crime at places: a longitudinal study of street segments in the city of Seattle. *Criminology*. 2004;42:283-320.
23. Nagin DS, Odgers CL. Group-based trajectory modeling in clinical research. *Annu Rev Clin Psychol*. 2010;6:109-138.
24. Nagin D, Piquero A. Using the group-based trajectory model to study crime over the life course. *J Crim Just Educ*. 2010;21(2):105-116.
25. Melhuish E, Sylva K, Sammons P, Siraj-Blatchford I, Taggart B. *Effective Pre-School, Primary & Secondary Education Project: an Investigation of Children's Learning Trajectories from 3 to 11 Years of Age in Literacy and Numeracy*. Institute of Education, University of London; 2011.
26. Sutcliffe AG, Gardiner J, Melhuish E. Educational Progress of looked-after children in England: a study using group trajectory analysis. *Pediatrics*. 2017;140(3):e20170503.
27. Shearer DM, Thomson WM, Broadbent JM, McLean R, Poulton R, Mann J. High-risk glycated hemoglobin trajectories established by mid-20s: findings from a birth cohort study. *Br Med J Open Diabet Res Care*. 2016;4:e000243.
28. Broadbent JM, Thomson WM, Poulton R. Trajectory patterns of dental caries experience in the permanent dentition to the fourth decade of life. *J Dent Res*. 2008;87(1):69-72.
29. Thomson WM, Shearer DM, Broadbent JM, Foster Page LA, Poulton R. The natural history of periodontal attachment loss during the third and fourth decades of life. *J Clin Periodontol*. 2013;40(7):672-680.
30. Kline P. *A Pyschometrics Primer*. Free Association Books; 2000.
31. Schuwirth LWT, van der Vleuten CPM. How to design a useful test: the principles of assessment. In: Swanwick T, ed. *Understanding Medical Education: Evidence, Theory and Practice*. 2nd ed. John Wiley & Sons, Ltd.; 2014.
32. Jolly B, Dalton MJ. Written assessment. In: Swanwick T, Forrest K, O'Brien BC, eds. *Understanding Medical Education: Evidence, Theory, and Practice*. 3rd ed. Wiley-Blackwell; 2019.
33. Boursicot KAM, Roberts TE, Burdick WP. Structured assessments of clinical competence. In: Swanwick T, Forrest K, O'Brien BC, eds. *Understanding Medical Education: Evidence, Theory, and Practice*. Wiley-Blackwell; 2019.
34. Nagin D. *Group-Based Modeling of Development*. Harvard University Press; 2005.
35. Webb GI. Posterior probability. In: Sammut C, Webb GI, eds. *Encyclopedia of Machine Learning*. Springer; 2011.
36. Raftery AE. Bayesian model selection in social research. *Sociol Methodol*. 1995;25:111-163.
37. Hoffman JIE. Hypothesis testing: the null hypothesis, significance and type I error. *Basic Biostatistics for Medical and Biomedical Practitioners*. 2nd ed. Elsevier Science; 2019:159-171.
38. Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika*. 1951;16:297-334.
39. Holmboe HS, Hawkins RE, Huot SJ. Effects of training in direct observation of medical residents' clinical competence: a randomized trial. *Ann Intern Med*. 2004;140(11):874-881.
40. Brown G, Manogue M, Martin M. The validity and reliability of an OSCE in dentistry. *Eur J Dental Educ*. 1999;3(3):117-125.
41. Hodges B. Validity and the OSCE. *Med Teach*. 2003;25(3):250-254.
42. Park RS, Chibnall JT, Blaskiewicz RJ, Furman GE, Powell JK, Mohr CJ. Construct validity of an objective structured clinical examination (OSCE) in psychiatry: associations with the clinical skills examination and other indicators. *Acad Psychiatry*. 2004;28(2):122-128.

43. Varkey P, Natt N, Lesnick T, Downing S, Yudkowsky R. Validity evidence for an OSCE to assess competency in systems-based practice and practice-based learning and improvement: a preliminary investigation. *Acad Med*. 2008;83(8):775-780.
44. Taghva A, Panaghi L, Rasoulian M, Bolhari J, Zarghami M, Esfahani MN. Evaluation of reliability and validity of the psychiatry OSCE in Iran. *Acad Psychiatry*. 2010;34(2):154-157.
45. Barry M, Bradshaw C, Noonan M. Improving the content and face validity of OSCE assessment marking criteria on an undergraduate midwifery programme: a quality initiative. *Nurse Educ Pract*. 2013;13(5):477-480.
46. Nickbakht M, Amiri M, Latifi SM. Study of the reliability and validity of objective structured clinical examination (OSCE) in the assessment of clinical skills of audiology students. *Global J Health Sci*. 2013;5(3):64-68.
47. Sam AH, Westacott R, Gurnell M, Wilson R, Meeran K, Brown C. Comparing single-best-answer and very-short-answer questions for the assessment of applied medical knowledge in 20 UK medical schools: cross-sectional study. *BMJ Open*. 2019;9(9):e032550.
48. Yepes-Rios M, Dudek N, Duboyce R, Curtis J, Allard RJ, Varpio L. The failure to fail underperforming trainees in health professions education: a BEME systematic review: BEME guide No. 42. *Med Teach*. 2016;38:1092-1099.
49. Dudek NL, Marks MB, Regehr G. Failure to fail: the perspectives of clinical supervisors. *Acad Med*. 2005;80:S84-S87.
50. Cleland JA, Knight LV, Rees CE, Tracey S, Bond CM. Is it me or is it them? Factors that influence the passing of underperforming students. *Med Educ*. 2008;42:800-809.
51. Chambers DW. Toward a competency-based curriculum. *J Dent Educ*. 1993;57:790-793.
52. Chambers DW. Competency-based dental education in context. *Eur J Dent Educ*. 1998;2:8-13.
53. Licari FW, Chambers DW. Some paradoxes in competency-based dental education. *J Dent Educ*. 2008;72:8-18.
54. Bush HM, Schreiber RS, Oliver SJ. Failing to fail: clinicians' experience of assessing underperforming dental students. *Eur J Dent Educ*. 2013;17:198-207.
55. Wilby KJ, Dolmans DHJM, Austin Z, Govaerts MJB. Assessors' interpretations of narrative data on communication skills in a summative OSCE. *Med Educ*. 2019;53:1003-1012.
56. Berendonk C, Stalmeijer RE, Schuwirth LWT. Expertise in performance assessment: assessors' perspectives. *Adv Health Sci Educ Theory Pract*. 2013;18:559-571.
57. Gingerich A, Regehr G, Eva KW. Rater-based assessments as social judgments: rethinking the etiology of rater errors. *Acad Med*. 2011;86:S1-S7.
58. Cleland J, Arnold R, Chesser A. Failing finals is often a surprise for the student but not the teacher: identifying difficulties and supporting students with academic difficulties. *Med Teach*. 2005;27:504-508.
59. Eva KW, Reiter HI, Trinh K, Wasi P, Rosenfeld J, Norman GR. Predictive validity of the multiple mini-interview for selecting medical trainees. *Med Educ*. 2009;43:767-775.
60. McManus IC, Thompson M, Mollon J. Assessment of examiner leniency and stringency ('hawk-dove effect') in the MRCP(UK) clinical examination (PACES) using multi-facet Rasch modelling. *BMC Med Educ*. 2006;6:42.
61. Hodges B. Assessment in the post-psychometric era: learning to love the subjective and collective. *Med Teach*. 2013;35:564-568.
62. Lockyer J, Carraccio C, Chan M, et al. Core principles of assessment in competency-based medical education. *Med Teach*. 2017;39:609-616.
63. Hoang NS, Lau JN. A call for mixed methods in competency-based medical education: how we can prevent the overfitting of curriculum and assessment. *Acad Med*. 2018;93:996-1001.
64. Downing SM. Validity: on the meaningful interpretation of assessment data. *Med Educ*. 2003;37:830-837.
65. Harris P, Bhanji F, Topps M, et al. Evolving concepts of assessment in a competency-based world. *Med Teach*. 2017;39:603-608.
66. Royal KD. Four tenets of modern validity theory for medical education assessment and evaluation. *Adv Med Educ Pract*. 2017;8:567-570.
67. Roeder K, Lynch KG, Nagin DS. Modeling uncertainty in latent class membership: a case study in criminology. *J Am Stat Assoc*. 1999;94:766-776.
68. Nagin DS, Tremblay RE. Developmental trajectory groups: fact or a useful statistical fiction? *Criminology (Beverly Hills)*. 2005;43:873-904.
69. NHS Education for Scotland. *How to Assess your Vocational Dental Practitioner (VDP) Using the Longitudinal Evaluation of Performance (LEP)*. NHS Education for Scotland; 2021.

How to cite this article: Dickie J, Sherriff A, McEwan M, Bell A, Naudi K. Longitudinal assessment of undergraduate dental students: Building evidence for validity. *Eur J Dent Educ*. 2023;27:1136-1150. doi:10.1111/eje.12908