International Journal of Osteopathic Medicine 31 (2019) 28–35

Contents lists available at ScienceDirect

International Journal of Osteopathic Medicine

journal homepage: www.elsevier.com/locate/ijosm

Original Article

Evaluation of teaching in a student-led clinic environment: Assessing the reliability of a questionnaire

Brett Vaughan*

College of Health & Biomedicine, Victoria University, Melbourne, Australia

ARTICLE INFO

Keywords:
Educational measurement
Generalizability
validity
Osteopathic medicine
Medical education

ABSTRACT

Introduction: The Osteopathy Clinical Teaching Questionnaire (OCTQ) has been designed to evaluate the quality of clinical teaching in osteopathy student-led clinics. Previous research has provided evidence for the scoring, generalisation and implications components of the validity argument for the OCTQ. The aim of the present study was to assess the reliability of the OCTQ providing further evidence for the validity argument.

Method: Senior students in the final years of two Australian osteopathy programs completed the OCTQ for each clinical educator with whom they had worked with during a semester. Generalisability analysis, test-retest reliability and internal structure estimation were used to investigate the reliability.

Results: Each of the forty-one clinical educators received an average of 5.97 evaluations resulting in a G-coefficient of 0.63. D-studies demonstrated eight evaluations are required for a coefficient of 0.7 and fourteen for a coefficient of 0.8. Test-retest reliability demonstrated substantial to almost perfect agreement for all but one OCTQ item. Internal structure estimations using Cronbach's alpha and McDonald's omega were both 0.93.

Conclusion: The results suggest that the OCTQ is a reliable tool to provide feedback to clinical educators, and potentially, used to inform decisions to reward clinical educators for their performance.

Introduction

The Osteopathy Clinical Teaching Questionnaire (OCTQ) has been developed as a student evaluation of their learning experience with an individual clinical teacher during student-led clinics in the on-campus university clinic environment. Like other evaluations of clinical teaching [1,2], the OCTQ explores the students' opinion of the quality of teaching provided by their clinical educator(s). Work undertaken thus far to develop the OCTQ has been guided by current recommendations with respect to exploration of the questionnaire measurement properties, including the use of classical test [3] and item response theory approaches [4]. The current paper reports on an extension of the development work - to explore the reliability of the OCTQ.

Student evaluations of teaching such as the OCTQ are often used to inform faculty and professional development activities, and potentially curricula changes. There are also numerous examples of where these are used for employment, tenure and promotion decisions by program administrators [2,5]. Where they are used for such decisions, the validity of interpretation of the score derived from the evaluation(s) needs to be established or argued [2]. This argument allows for defensible decision-making and potentially greater engagement with, and trust in, the evaluation process by students, clinical educators, and program administrators. The reliability of a questionnaire also contributes to trustworthy and defensible decision-making when combined with other evidence to develop a validity argument.

The reliability of a questionnaire contributes valuable information to the overall understanding of its measurement properties. Providing such evidence allows stakeholders to make a judgement with respect to the dependability of the scores derived from the questionnaire. Multiple forms of reliability have been reported in the clinical teaching evaluation literature, including internal consistency (typically calculated as Cronbach's alpha) in addition to inter-rater reliability, and test-retest reliability [2,6,7]. More recently, generalisability theory has been applied in the development of clinical teaching evaluation questionnaires [8–15]. However, few studies utilise multiple approaches to the exploration of reliability to support the validity argument [16]. Each reliability statistic has its own limitation, primarily related to the data-driven nature of the calculation of these statistics. Therefore, there is great value in using multiple forms of reliability calculation as this will ensure "… ample evidence be [is] accumulated to establish the reliability of scores before using an instrument in practice" [16].

Kane [17,18] has provided numerous commentaries on the use of an argument-based approach to validity. Kane's framework is predicated

* Brett Vaughan College of Health & Biomedicine, City Flinders Lane Campus, Victoria University, PO Box 14428, Melbourne, VIC, 8001, Australia.
E-mail address: brett.vaughan@vu.edu.au.

https://doi.org/10.1016/j.ijosm.2018.11.001
Received 20 August 2018; Accepted 2 November 2018

on the idea that multiple sources of evidence are required in order to demonstrate validity – one develops an argument for the validity of the interpretation of a score. The premise is that an assessment or evaluation tool in itself cannot be valid, it is the interpretation of the score derived from the tool that may demonstrate validity. Kane outlined a sequential approach to developing and presenting relevant evidence for the validity argument: *scoring*, *generalisation*, *extrapolation* and *implications*. Importantly Kane's approach does not dictate the type of evidence that can be presented for each part of the argument. This allows for research into different contexts to contribute evidence to building an argument, and the flexibility to utilise qualitative, quantitative and mixed methods research designs as part of this approach.

Evidence for the validity argument for the Osteopathy Clinical Teaching Questionnaire (OCTQ) has been presented previously [3,4]. This evidence to date provides support for the *scoring*, *generalisation* and *implication* arguments for the interpretation of the score derived from the OCTQ however the *extrapolation* argument requires further development. A key component of the validity argument that requires addressing from the outset is the proposed interpretation of the score derived from the questionnaire. In the context of the OCTQ, the proposed interpretation is the *quality of clinical teaching provided by a clinical educator in the osteopathy student-led, on-campus clinic environment*. The aim of the current study is to present a suite of reliability statistics to evaluate the generalisability, retest reliability and criterion validity of the Osteopathy Clinical Teaching Questionnaire (OCTQ). This information provides further evidence for the validity argument for the interpretation of the score derived from the OCTQ.

Method

This study was approved by the Victoria University and Southern Cross University Human Research Ethics Committees (HRE15-238 and ECN15-301).

Participants

Students enrolled in year 4 and year 5 of the osteopathy programs at Victoria University (VU) (Melbourne, Australia) and Southern Cross University (SCU) (Lismore, Australia) were invited to complete a paper version of the Osteopathy Clinical Teaching Questionnaire during October and November 2015.

Questionnaire

The Osteopathy Clinical Teaching Questionnaire (OCTQ) [3,4] is a 15-item measure of clinical teaching quality in osteopathy student-led, on-campus clinics – 12 items measuring clinical teaching quality and 3 global rating items. Each of the 12 clinical teaching quality items are scored on a 1 (strongly disagree) to 5 (strongly agree) Likert-type scale and takes approximately 2 min to complete. The reliability estimates for the OCTQ have been calculated as greater than 0.85 and the calculation of a total score (not including the 3 global items) is a sufficient statistic [4]. When calculating the total score, item 12 requires rescoring - the *strongly disagree/disagree* options are collapsed to a 1 with the other three options scored as 2, 3 and 4 respectively. The total available score for the 12 items on the OCTQ is 59 and can be converted to an interval-level score using the table at Appendix 1. Each of the three global items are rated on their own five-point Likert-type scale.

Data collection

At both Victoria University and Southern Cross University, students were asked to complete one questionnaire for each of the osteopathy clinical educators they had worked with in the on-campus, student-led clinics during the second-half of the 2015 academic year. When completing the OCTQ, the student was asked to identify the clinical educator being rated and their own gender however they were not required to personally identify themselves and could complete the questionnaires at a time of their choosing. Completion of an evaluation was not a requirement to receive a grade or satisfactorily complete a clinical unit in either osteopathy program.

At Victoria University, in order to evaluate the test-retest reliability of the OCTQ, students in year 4 only were asked to write a self-generated code on each questionnaire they completed at both time 1 (T1) and time 2 (T2) in order to match responses. Students completed the OCTQ at T1 then undertook their normal one-and-a-half-hour practical skills class. This process was used to ensure that the students were not exposed to any clinical teaching in between the administrations of the questionnaire. At the end of the class (T2), the students were asked to complete the OCTQ again for each educator they rated at T1.

Data analysis

Data were entered into SPSS (version 21) to organise the export to other programs for the data analysis. Item 12 was rescored as per Vaughan [4] where the *strongly disagree* and *disagree* options were collapsed into one. Data were then exported to *R* [19] to generate the descriptive statistics using the *psych* package [20].

Generalisability analysis

Background. Generalisability theory (GT) is based on an extension of the analysis of variance (ANOVA) in classical test theory and is a useful analysis where there are multiple sources of variance (i.e. students, examiners, items, cases, patients) in the total score derived from a measure [21]. The measurement score is decomposed into these multiple sources, or 'facets' in GT, allowing for the evaluation of how the individual facet and its interactions with other facets, contribute to the score variance. The *object of the measurement* [21] or *facet of differentiation* [22] is the entity being measured and the score for this entity is thought to represent the universe score [23]. The other entities are termed either fixed or random *facets of generalization*. A fixed facet is one that, for the purposes of the measurement, does not change (i.e. number of items on the measure) [23]. Conversely, a random facet is one where the number can change as these facets are typically drawn from a universe of entities (i.e. raters, examiners, students). In GT, consideration is also given to whether a facet is crossed or nested within another facet [21]. A facet is crossed with another where each level of the facet interacts with the other (i.e. all students are assessed by all examiners). Conversely a facet is nested when one level of the facet interacts with a portion of the other facet (i.e. some students are assessed by one examiner, and other students by another examiner).

Together this information is used to design the generalisability analysis. Bloch and Norman [21] summarise the desired outcome of a generalisability analysis as

> "… to what extent can we extrapolate the results achieved on a limited sample of test tasks, measured under unique test conditions to a universe of tasks and conditions, from which the specific test set has been drawn more or less arbitrarily".

The generalisability analysis produces variance component values and two reliability coefficients, $E\rho^2$ (relative error) and Φ (absolute error). The variance components can be converted to percentages by dividing them by the total score variance in order to assist in their interpretation [23]. The decision about which coefficient to report is based on whether the desire is to make a comparison with other individuals within the *facet of generalization* ($E\rho^2$) (i.e. normative decision making) or the interest is only in the individual (Φ) [21,23].

The previous is a description of a G study – the derivation of a reliability coefficient by estimating the variance attributable to different facets [23]. GT allows for the variation of the *facets of generalization* to model the impact of changing the number or level of the facet on the reliability coefficient. This is referred to as a decision (D) study and is

designed to develop "… a measurement that minimizes error for a particular purpose" [24]. A more in-depth review of GT, its mathematical basis and the practical applications beyond that described in the present study is provided by Bloch and Norman [21].

Analysis. A generalisability analysis was performed to determine the variance components and reliability of the OCTQ and ascertain how many questionnaires are required for a reliable decision. The design of the present study was i x (e:s) where 'i' represents the 12 OCTQ items, 'e' represents the clinical educators and 's' represents the students. This design is consistent with other G studies of clinical teaching evaluations [8,12–15], and is succinctly described by Bloch and Norman [21]:

> "Teacher [clinical educator] is the facet of differentiation, and here student (rater), a facet of generalization, is nested in teacher since each student belongs to only one section. But item (on the scale) is crossed with teacher, since all teachers are rated on the same items"

The design was unbalanced in that each clinical educator received a different number of ratings from the students. The Φ coefficient was reported as the performance of an individual clinical educator was not compared to that of their peers (an *absolute* rather than *relative* decision in GT terminology).

Using the variance components, a D study was performed to determine how many questionnaires would be required to achieve a generalisability coefficient of 0.80 [23]. This value is widely considered to be acceptable for high-stakes decision making. Data were imported into *G_string IV* [25] for the generalisability analysis (G-study) and decision study (D-study). The number of *Items* was fixed at 12 as these were the only items of interest in the present study [7,13].

Temporal stability (test-retest reliability)

Temporal stability was evaluated using weighted kappa in *R* [19] using the *psych* package [20]. Norman and Streiner [26] have previously reported that weighted kappa is equivalent to an intraclass correlation coefficient ($ICC_{2,1}$) and can be used interchangeably. Weighted kappa was chosen as the data were ordinal in nature. The interpretation of the weighted kappa was based on Landis and Koch [27]: < 0 *Less than chance agreement*; 0.01–0.20 *Slight agreement*; 0.21–0.40 *Fair agreement*; 0.41–0.60 *Moderate agreement*; 0.61–0.80 *Substantial agreement*; and 0.81–0.99 Almost *perfect agreement*.

Relationship between global ratings and total scores

The relationship between the three global ratings and total Rasch-converted OCTQ score was evaluated using Spearman's rho (ρ) and interpreted according to Hinkle, Wiersma [28]: 0–0.30 (negligible); 0.30–0.50 (low); 0.50–0.70 (moderate); 0.70–0.90 (high); 0.90–1.00 (very high). The total Rasch-converted OCTQ score [4] was used as the data are interval-level rather than ordinal potentially allowing for greater sensitivity in the difference between total questionnaire scores.

Reliability estimation

Cronbach's alpha (α) and McDonald's omega total (ωt) (and their confidence intervals) were calculated as the reliability estimates using the *MBESS* package [29] in *R* [19]. Bootstrapping was applied to confirm the CI's using 1000, 5000 and 10000 iterations. Only the responses from T1 were used for the analysis and as the OCTQ has been shown to be unidimensional [4] calculating both reliability estimates using all 12 OCTQ items is appropriate. Item-total correlations and alpha if item deleted statistics were generated using the *psych* package [20]. Cronbach's alpha is a widely reported reliability estimate [30] however numerous authors have highlighted issues with its use as a measure of internal consistency [30–32]. That said, there is little agreement in the literature about which estimates to calculate. Dunn, Baguley [30] suggest that McDonald's omega (ω) is a suitable alternative particularly where the tau-equivalence assumption for α is violated.

The test-retest Cronbach's alpha was calculated [33] in order to evaluate the effects of transient error in the test-retest administration of the OCTQ. Effects of item recall on alpha is reported to be less for test-retest alpha than the test-retest correlation, and an underestimate of reliability [33]. Tau-equivalency is an assumption underlying the use of test-retest alpha however Green [33] suggests this assumption is reported to be largely met in measures that demonstrate unidimensionality. Previous research has provided evidence for the unidimensionality of the OCTQ [4] therefore use of Cronbach's alpha is acceptable.

McDonald's omega hierarchical (ωh) provides an indication as to the reliable variance attributable to the general factor or latent construct being measured. Values greater than 0.50 [34] and 0.70 [35] have been suggested as acceptable. ωh was calculated using the *psych* package [20] in *R* [19].

Table 1
Overview of clinical educator and student responses.

		Total	Institution	
			Victoria University	Southern Cross University
Total responses		304	199 (65.4%)	105 (34.6%)
Responses by student gender	Male	156	96 (48.2%)	60 (57.1%)
	Female	136	95 (47.7%)	41 (39.0%)
Responses by clinical educator gender	Male	196	117 (58.8%)	79 (75.2%)
	Female	107	81 (40.7%)	26 (24.8%)

Note: some participants did not indicate either their gender or the gender of the clinical educator being rated.

Results

Three-hundred and four questionnaires were received evaluating 41 of the 44 clinical educators across the two institutions (Table 1). Descriptive statistics for the OCTQ are presented in Table 2.

A G-coefficient of 0.63 was obtained based on an average of 5.97 questionnaires per clinical educator (range - 2 to 15 questionnaires per educator). The variance components are presented in Table 3 and the D-study coefficients are presented in Fig. 1. To provide feedback to an individual clinical educator that is reliable, 8 questionnaires are required to achieve a Φ coefficient of 0.70.

Test-retest reliability results are presented in Table 1. *Substantial* to *almost perfect* agreement was observed for all but item 12 where *fair* agreement was observed. *Almost perfect* agreement was observed for all three global ratings. Test-retest alpha was 0.90. These results suggest the OCTQ items, except for item 12, are stable across a short-term administration. The relationships between the total Rasch-converted OCTQ score and global rating one (I would do more clinics with this Clinical Educator) and two (I would recommend other students to work with this Clinical Educator) were both *moderate* ($\rho = 0.68$, $p < 0.01$). A *high* correlation was observed between the OCTQ total score and global rating three (Rate the overall effectiveness of this Clinical Educator as an educator/supervisor) ($\rho = 0.72$, $p < 0.01$).

Cronbach's α and McDonald's ωt were 0.93 [95%CI 0.91–0.94] and 0.93 [95%CI 0.92–0.94] respectively. These values suggest that approximately 7% of the variance in the OCTQ score is measurement error, and these values did not change in relation to the number of bootstrapping iterations. Item-total correlations were *moderate* or greater (range 0.63–0.83), the inter-item correlation average was 0.51 (range 0.50–0.53) (Table 1), and alpha did not improve if any OCTQ item was deleted. ωh was acceptable at 0.73 and suggests that 73% variance in the total OCTQ score is accounted for by the latent construct of clinical teaching quality in osteopathy student-led, on-campus clinics.

Table 2
Descriptive, intra-examiner and internal structure correlation statistics for the Osteopathy Clinical Teaching Questionnaire.

Item	Mean	St Dev	Median	Minimum	Maximum	Weighted Kappa	95% Confidence Interval	Item-total Correlation	Average Inter-item Correlation
1. Maintained a positive attitude towards me	4.64	0.66	5	1	5	0.86	0.77–0.96	0.70	0.52
2. Demonstrated humanistic attitudes in relating to patients (integrity, compassion and respect)	4.70	0.57	5	2	5	0.78	0.66–0.89	0.64	0.52
3. Showed genuine concern for my professional well-being	4.61	0.65	5	2	5	0.74	0.59–0.89	0.75	0.51
4. Has good communication skills	4.53	0.71	5	2	5	0.72	0.58–0.86	0.81	0.50
5. Is open to student questions and alternative approaches to patient management	4.46	0.81	5	1	5	0.83	0.74–0.92	0.72	0.52
6. Adjusted teaching to my needs (experience, competence, interest)	4.36	0.79	5	1	5	0.81	0.73–0.90	0.83	0.50
7. Promoted reflection on clinical practice	4.38	0.76	5	1	5	0.73	0.59–0.86	0.74	0.51
8. Emphasises a problem-solving approach rather than solutions	4.45	0.76	5	2	5	0.71	0.60–0.82	0.77	0.51
9. Asked questions to enhance my learning	4.45	0.80	5	1	5	0.75	0.64–0.86	0.75	0.51
10. Stimulates me to learn independently	4.39	0.84	5	1	5	0.80	0.69–0.91	0.80	0.50
11. Offered me suggestions for improvement when required	4.55	0.71	5	1	5	0.72	0.60–0.84	0.75	0.51
12. Demonstrated osteopathic, clinical examination and rehabilitation knowledge and skill(s)	3.70	0.55	4	2	4	0.50	0.26–0.75	0.63	0.53
Total score	53.29	6.42	55	28	59				
G1. I would do more clinics with this Clinical Educator	4.57	0.78	5	1	5	0.88	0.80–0.95		
G2. I would recommend other students to work with this Clinical Educator	4.59	0.74		1	5	0.95	0.90–0.99		
G3. Rate the overall effectiveness of this Clinical Educator as an educator/supervisor	4.49	0.76	5	1	5	0.97	0.94–1.00		

Table 3
Variance components for the generalisability study of the Osteopathy Clinical Teaching Questionnaire.

Effect	Degrees of freedom	Sum of Squares	Variance Component	Percentage Variance Component
Educator	39	327.78	0.06	10.5
Student:Educator	256	688.36	0.20	35.1
Items	11	214.99	0.06	10.5
Educator x Items	429	165.35	0.02	3.6
Items x Student:Educator[a]	2816	652.73	0.23	40.3

[a] Residual error.

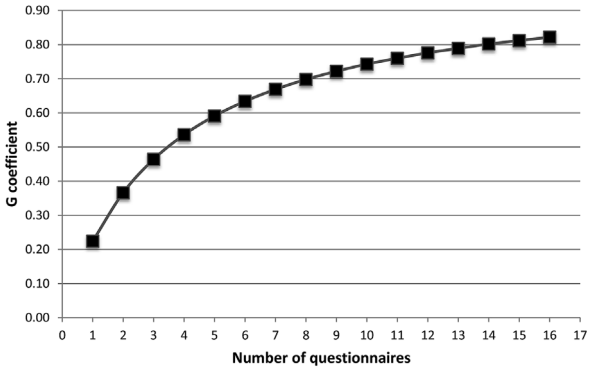

Fig. 1. Number of Osteopathy Clinical Teaching Questionnaires for different reliability coefficients.

Discussion

The present study extends on previous work by evaluating the reliability of the OCTQ to support trustworthy decision-making based on the results of the questionnaire. That is, program administrators and clinical educators can have a sense of 'trust' in the data based on the evidence supporting the validity of the score derived from the questionnaire described here.

Generalisability analysis

The G-study demonstrated that an average of 5.97 questionnaires achieved a Φ coefficient of 0.63. This is consistent with the initial analysis of a 9-item clinical teaching behaviours questionnaire by Keely, Oppenheimer [15]. A D-study using the OCTQ data suggests that questionnaires completed by eight different students on a single clinical educator provides a Φ coefficient of 0.7. This reliability coefficient provides evidence to support the use of the OCTQ as a feedback tool [36]. The information and scores could be used to guide program administrators about faculty development activities targeted to an individual clinical educator for example. In order to utilise the scores derived from the questionnaire for higher-stakes decisions such as promotion, tenure or teaching awards, a single clinical educator would require 14 individual student questionnaires. This number of questionnaires provides a Φ coefficient of 0.8, and is consistent with other clinical teaching evaluations published by Keely, Oppenheimer [15] and Hindman, Dexter [14]. Beyond this number of evaluations the reliability would not substantially increase (Fig. 1).

Calculation of the variance components that underpin these Φ coefficients provides further information about the psychometric properties of the OCTQ. There are no universal guidelines by which to determine the acceptability of a variance component value [23] and they need to be interpreted within the context of the individual study. This can also be a limitation applied to any study of reliability as the calculated coefficients may change, albeit minimally. When combined

with other reliability data, as in the present study, the impact of this limitation is reduced but does not negate the need for other researchers to evaluate reliability in their own clinical education environment.

The decision to report the Φ rather than the Ep^2 coefficient is supported by the small variance component for Educator (10.5%). The variance attributable to the differences between individual clinical educator OCTQ scores is approximately one-tenth [37]. The clinical educators in the present study are likely to already be performing at a high level, as evidenced by the high mean OCTQ score. Therefore, there is little difference between the mean scores for each Educator evaluated in the present study therefore there is little ability to rank the clinical educators based on their total OCTQ scores. Further comparisons between educators are not appropriate, and from a practical standpoint, would represent a normative decision that is unlikely to be defensible. de Oliveira Filho, Dal Mago [12] and Haider, Johnson [13], using the same G-study design as the present study, reported variance components of 56% and 28% respectively for the instructor/teacher [Educator] in their studies. The substantial difference in the variance component is likely due to the fact that de Oliveira Filho, Dal Mago [12] utilised a balanced design, whereas an unbalanced design that is more reflective of the clinical education environment was used in the present study.

The Student:Educator interaction in the present study accounts for the second largest variance (35.1%) and is consistent with other studies [12,14,38]. This result suggests that each student has a different perception of the quality of teaching provided by the clinical educator [12,15,23]. de Oliveira Filho, Dal Mago [12] also suggest that this interaction term could reflect differences in the students' interpretation of the questionnaire items, real differences in teaching quality between clinical educators, and/or a combination of all of these factors. Haider, Johnson [13] demonstrated a 4% variance component for this interaction in their study. These authors argued that stringency (all raters agreed on the same level for the teaching performance) may have accounted for this small variance component. The substantial difference between the variance components for Student:Educator in the present study and those presented by Haider, Johnson [13] provides support for the argument that students are rating each clinical educator differently, whilst using the same items on the OCTQ. This appears to be a strength of the twelve items that comprise the OCTQ.

It is not possible to deconstruct this Student:Educator variance component further using the current design, however it does provide an avenue for future research. The substantial contribution of this interaction to the score variance supports the need to obtain a suitable number of responses to make a trustworthy *absolute* decision, and avoid making *relative* decisions [39].

The Item facet provides an indication as the internal consistency of the measure used in the study [23]. The variance accounted for by Item was 10.5% suggesting the OCTQ items generated relatively similar mean scores (Table 1) with some variation in the level of difficulty of each item [14]. Vaughan [4] demonstrated that the OCTQ items evaluate different levels of clinical educator performance therefore variation in item difficulty is expected. The Item variance component is relatively consistent with the other reliability estimation methods in the present study (α and ωt = 0.93), and those described by Copeland and Hewson [8] (11%) and Haider, Johnson [13] (18%) using alternative questionnaires to evaluate clinical teaching.

The Educator x Item term accounted for 3.6% of the variance, and is somewhat less than the 23% demonstrated by Copeland and Hewson [8] and 16% by Haider, Johnson [13] but more consistent with Hindman, Dexter [14]. Again, this difference could be attributed to the respective study designs. Although these studies treated Student as a random factor, Copeland and Hewson [8] and Haider, Johnson [13] used 5 randomly selected responses whereas the present study utilised all responses, regardless of how many a clinical educator received. Given the random nature of the selected responses, the relative ranking of the clinical educators would change if different students rated the educators in the studies by Copeland and Hewson [8] and Haider, Johnson [13]. In contrast, there would be very little change in the relative rankings in the present study with the small Educator x Item variance component.

Items crossed with Student nested in Educator reflects both random and systematic error variance [10]. This was the largest contribution to the variance of the total OCTQ score at just over 40%, and is consistent with studies by Copeland and Hewson [8] (46%), Hindman, Dexter [14] (43%), and Haider, Johnson [13] (36%). Because of the nested design of the G-study it is not possible to deconstruct this variance component further [10]. This result does however provide some support for the notion that the students may be interpreting the OCTQ items differently [8,13] - a 3-way interaction between the student-clinical educator-OCTQ item [14]. As highlighted previously, this difference in interpretation could be due to an actual difference in the interpretation of the items between students, or the influence of other factors (e.g. age, level of training, clinical learning environment, personality) on item interpretation [10].

A number of studies have investigated the inter-rater reliability of clinical teaching evaluations [12,40–42]. Such studies make the implicit assumption that different students will have the same (or similar) perception of a clinical educator's teaching quality. The G-study presented here demonstrated an inter-rater reliability coefficient of 0.63 suggesting that approximately 6 OCTQ responses would provide a moderate level of inter-rater reliability. Students in the present study did not appear to have the same perception of a single clinical educator. It is known that, amongst other factors, the interpersonal relationship between the clinical educator and an individual student [10,14,43], and the clinical learning environment [44] influence responses to clinical teaching evaluations. These factors may account for the moderate inter-rater reliability. Moderate inter-rater reliability may also be a valuable property of clinical educator evaluations. If all students are consistent in their evaluation of a single clinical educator the utility of the information that can be derived from the questionnaire becomes somewhat limited.

Temporal stability

To ascertain an accurate indication of temporal stability of the OCTQ items, weighted Kappa was used in a subset of the student population reported here. Temporal stability of OCTQ items 1 to 11 was *substantial to almost perfect*. Importantly, no student was exposed to their clinical educator in the period during the data collection. Therefore, it is not possible for any interaction between educator and student to influence these results. Given students were required to rate multiple clinical educators they are less likely to remember the previous rating, reinforcing the temporal stability results. The high test-retest alpha (0.90) and narrow 95% confidence intervals for each of these items further supports the accuracy of these results.

Item 12 was the only item that did not demonstrate substantial to almost perfect agreement. This item evaluates the students' perception of demonstration of skills related to manual therapy practice. Previous work to develop the OCTQ demonstrated that this item was the only one that required rescoring [4]. The *strongly disagree* and *disagree* options were collapsed together to create the OCTQ total score. Whilst this item did demonstrate fit to the Rasch model in a previous study [4], these two issues (item rescoring and moderate agreement) suggest that the item may require further investigation.

A possible reason could be that suggested by Mackillop, Parker-Swift [45] in their study. These authors identified that compound items on a multisource feedback form were less reliable than non-compound items. Item 12 *Demonstrated osteopathic, clinical examination and rehabilitation knowledge and skill(s)* could be classified as a compound item in that it asks the respondent to evaluate the clinical educator on multiple aspects of their performance within the one item. Splitting item 12 may assist in improving its temporal stability. Future

work could investigate two options for the item:

1 Split the item into Demonstrated osteopathic, clinical examination and rehabilitation knowledge and Demonstrated osteopathic, clinical examination and rehabilitation skills; or
2 Split the item into Demonstrated osteopathic knowledge and skills, Demonstrated clinical examination knowledge and skills, and Demonstrated rehabilitation knowledge and skills.

The cognitive load associated with this item may be too high for the student when evaluating their clinical educator. That is, they are being asked to evaluate the clinical educator on multiple areas – the compound item. Adding to the increased cognitive load is the student rating multiple clinical educators in a single administration in the present study. Whether the temporal stability would improve if the student was asked to repeatedly evaluate only one clinical educator requires further investigation.

Relationship with global ratings

Bierer and Hull [46] suggest that clinical teaching evaluations that are to be used for summative decisions should contain a global rating item, and it is posited here that this should also apply to formative decisions. The OCTQ has three global rating items: one to capture a students' overall perception of effectiveness; and two 'satisfaction' items.

Rate the overall effectiveness of this Clinical Educator as an educator/supervisor demonstrated a high correlation with the total Rasch-converted OCTQ score ($\rho = 0.72$) and provides some evidence for criterion validity [11,47]. This high correlation provides support for the argument that the total OCTQ score captures a substantial proportion of the effectiveness of the teaching and supervision provided by the clinical educator [11,48] and is within the 0.40–0.80 range suggested by van der Leeuw, Lombarts [47].

Satisfaction could provide an indicator of quality processes [49]. *I would do more clinics with this Clinical Educator* and *I would recommend other students to work with this Clinical Educator* are the two global satisfaction items on the OCTQ. Both were *moderate* correlations but only slightly less than the effectiveness global rating described above. The correlations suggest that the interpersonal and teaching domains covered by the OCTQ [4] relate to satisfaction with the performance of the clinical educator. However, given the *moderate* correlations, it is likely there are other factors beyond that captured by the OCTQ that influence satisfaction with an educators' performance. Examples of these factors could be level of supervision provided [50], clinical learning environment, and the student-educator personality interaction [10,14,43].

Reliability estimations

Reliability estimates for the OCTQ were also calculated. Consistent with previous work [4] both α and ωt were both above 0.92. In the case of ωt the value of 0.93 indicates the general factor (latent construct) and group factors account for 93% of the variance in the total OCTQ score. The group factors have been described previously [4] and potentially represent the *interpersonal* and *clinical teaching* domains [6]. ωh provides an indication as to the variance attributable to the general factor (latent construct of quality of clinical teaching in osteopathy) [51]. In the present study, 73% of the total score variance is attributable to the general factor only. This is lower than the 86% observed in previous work on the OCTQ [4] but above an acceptable level [35]. This finding supports the previous assertion that reliability coefficients are data driven, will vary between datasets and require calculation with each administration. The current study is also the first to utilise McDonald's omega as a reliability estimation as part of the evidence for the validity of a clinical teaching evaluation, and other authors are encouraged to report the same when using the OCTQ or other clinical teaching evaluation tool.

Cronbach's α was also acceptable and removing an item did not improve the α score. Item-total correlations were also acceptable (0.63–0.81). The α value in the present study was over 0.90 suggesting item redundancy [52]. However removal of an item does not improve the α value, the item-total correlations were not high [15], and the OCTQ was constructed in such a way as to limit the inter-item relationships [4]. Therefore, item redundancy is unlikely to be an issue. Further, other authors of clinical teaching evaluations have demonstrated α values over 0.90 [7,8,12]. The α and ω statistics presented here provide further evidence to support the valid calculation of a total score for the OCTQ as it measures one latent construct – quality of clinical teaching in osteopathy student-led clinics [4].

Validity argument

Kane's validity argument approach [17] suggests evidence be provided for four aspects in order to argue for the validity of the interpretation of a score. Evidence for the *scoring* argument continues to be provided in the form of the reliability estimations. McDonald's omega hierarchal (ωh) [51] was above 0.70 in the present study and a previous study investigating the construct validity of the OCTQ [4]. When combined, both studies provide evidence to support the calculation of a total score for the OCTQ.

The statistical approach employed in the present study provides evidence for the *generalisation* argument. The students and clinical educators are drawn from the possible 'universe' [21,23] of stakeholders in the evaluation of clinical teaching in osteopathy student-led, on-campus clinics. This allows the results of the present study to be generalised to Australian osteopathy programs. Given the development of the OCTQ in a previous study [4] also included evaluations from osteopathy programs in New Zealand and the United Kingdom, it may be possible to argue for the *generalisation* to these programs. However, further evidence is required to support this assertion.

Reliable feedback can be achieved with eight individual students completing an evaluation on a single clinical educator. This information provides evidence for the implications argument in that the OCTQ can be used to provide feedback based on the student's perception of clinical educator performance. Temporal stability of the OCTQ items provides further support however the evidence is tempered by the *fair* intra-rater reliability for item 12. Further work is required to ensure the temporal stability of item 12 before it can be used for high-stakes decision-making – its use for guiding feedback and professional development is acceptable however.

Limitations

There are a number of limitations in the present study. Firstly, the study was undertaken in two Australian teaching institutions which may limit the generalisability of the results to other non-United States osteopathy teaching programs.

The anonymous nature of the responses to the OCTQ prevents a definitive statement about the response rates in the present study. However, there were approximately 130 students eligible to participate across both teaching programs at the time of the study. Therefore more than 300 responses provides an indication that the responses are representative of the 'universe' [23] from which the data could possibly be drawn at the time of the study. Each Clinical Educator also received six evaluations on average which represents the number of students the typical educator would supervise in an Australian osteopathy student-led, on-campus clinical setting [53].

Whilst the average number of questionnaires completed for each clinical educator was six, there was a range from two to fifteen. It is possible that a student may have only evaluated one or two clinical educators when they worked with more over the period of the study. This may have also led to the situation where the student chose to

evaluate only the clinical educators that they had a strong opinion about (whether positive or negative). Therefore, those clinical educators performing at the expected level may not have been evaluated.

The temporal stability of the OCTQ items could have been evaluated within the G-study. The G-study design would have included an Occasions facet representing the multiple administrations of the OCTQ [10]. This was not possible for logistical reasons as the ability to sample students between their clinic times was limited by timetabling. Such a study design is important to ensure that the students are not exposed to the clinical educators whom they are evaluating in the period between administrations [10].

Conclusion

The present study sought to investigate the reliability and number of questionnaires that would be required to be completed by students for a single clinical educator in two Australian osteopathy programs. This information can then be used to provide the clinical educator with feedback about their clinical teaching performance from the standpoint of the students. Eight questionnaires will provide a reliable indication as to the educator's performance. The information derived from these 8 questionnaires could be used to provide feedback to an Australian osteopathy clinical educator to assist them to identify their strengths, improve their performance and guide professional development. All but one OCTQ item, and the three global rating items demonstrated *substantial* to *almost perfect* agreement. This issue with the test-retest stability of one item requires further investigation prior to the OCTQ being used for promotion, tenure and teaching award decisions. The reliability estimations are consistent with previous work and support the validity of the calculation of a total score for the OCTQ. These results provide evidence for all four aspects of Kane's validity argument: scoring, generalisation, extrapolation and implications. Further work is now required to address the issue with one OCTQ item. Future studies will also explore the relationship of the OCTQ to clinical educator self-evaluations and self-efficacy, and how other clinical education environmental factors relate to OCTQ scores to continue building evidence for the validity argument.

Conflicts of interest

Brett Vaughan is a section editor for the International Journal of Osteopathic Medicine. He was not involved in review or editorial decisions regarding this manuscript.

Funding

This work was supported by an Australian Government Research Training Program Scholarship.

Ethical approval

This study was approved by the Victoria University and Southern Cross University Human Research Ethics Committees (HRE15-238 and ECN15-301).

Availability of data and material

The data set(s) supporting the results of this article are available in the *figshare* repository: 10.6084/m9.figshare.6667913.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijosm.2018.11.001.

References

[1] Beckman TJ, Cook DA, Mandrekar JN. What is the validity evidence for assessments of clinical teaching? J Gen Intern Med 2005;20(12):1159–64.
[2] Fluit C. Assessing the quality of clinical teachers. J Gen Intern Med 2010;25(12):1337–45.
[3] Vaughan B. Developing a clinical teaching quality questionnaire for use in a university osteopathic pre-registration teaching program. BMC Med Educ 2015;15(1):70.
[4] Vaughan B. Exploring the measurement properties of the osteopathy clinical teaching questionnaire using Rasch analysis. Chiropr Man Ther 2018;26(1):13.
[5] Oermann MH, Conklin JL, Rushton S, Bush MA, editors. Student evaluations of teaching (SET): guidelines for their use. Nurs Forum (Auckl). Wiley Online Library; 2018.
[6] Beckman TJ, Ghosh AK, Cook DA, Erwin PJ, Mandrekar JN. How reliable are assessments of clinical teaching? J Gen Intern Med 2004;19(9):971–7.
[7] Nation JG, Carmichael E, Fidler H, Violato C. The development of an instrument to assess clinical teaching with linkage to CanMEDS roles: a psychometric analysis. Med Teach 2011;33:e290–6.
[8] Copeland HL, Hewson MG. Developing and testing an instrument to measure the effectiveness of clinical teaching in an academic medical centre. Acad Med 2000;75(2):161–6.
[9] Der Hem-Stokroos V, Van der Vleuten C, Daelmans H, Haarman H, Scherpbier A. Reliability of the clinical teaching effectiveness instrument. Med Educ 2005;39(9):904–10.
[10] Conigliaro RL, Stratton TD. Assessing the quality of clinical teaching: a preliminary study. Med Educ 2010;44(4):379–86.
[11] Boerboom T, Dolmans D, Jaarsma A, Muijtjens A, Van Beukelen P, Scherpbier A. Exploring the validity and reliability of a questionnaire for evaluating veterinary clinical teachers' supervisory skills during clinical rotations. Med Teach 2011;33(2):e84–91.
[12] de Oliveira Filho GR, Dal Mago AJ, Garcia JHS, Goldschmidt R. An instrument designed for faculty supervision evaluation by anesthesia residents and its psychometric properties. Anesth Analg 2008;107:1316–22.
[13] Haider SI, Johnson N, Thistlethwaite JE, Fagan G, Bari MF. WATCH: warwick assessment insTrument for clinical teacHing: development and testing. Med Teach 2015;37(3):289–95.
[14] Hindman BJ, Dexter F, Kreiter CD, Wachtel RE. Determinants, associations, and psychometric properties of resident assessments of anesthesiologist operating room supervision. Anesth Analg 2013;116(6):1342–51.
[15] Keely E, Oppenheimer L, Woods T, Marks M. A teaching encounter card to evaluate clinical supervisors across clerkship rotations. Med Teach 2010;32(2):e96–100.
[16] Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. Am Med J 2006;119(2). 166.e7-.e16.
[17] Vaalidation Kane MT. Brennan R, editor. Educational measurement. fourth ed.Westport, USA: Praeger Publishers; 2006. p. 17–64.
[18] Kane MT. Validating the interpretations and uses of test scores. J Educ Meas 2013;50(1):1–73.
[19] R Core Team. R: A language and environment for statistical computing Vienna. Austria: R Foundation for Statistical Computing; 2015 Available from: https://www.R-project.org/, Version 3.2.2.
[20] Revelle W. Psych; procedures for personality and psychological research 1.5 eighth ed. Evanston, Illinois, USA: Northwestern University; 2015.
[21] Bloch R, Norman G. Generalizability theory for the perplexed: a practical introduction and guide: AMEE Guide No. 68. Med Teach 2012;34(11):960–92.
[22] Streiner DL, Norman GR, Cairney J. Health measurement scales: a practical guide to their development and use. fifth ed. Oxford University Press; 2014.
[23] Briesch AM, Swaminathan H, Welsh M, Chafouleas SM. Generalizability theory: a practical guide to study design, implementation, and interpretation. J Sch Psychol 2014;52(1):13–35.
[24] Shavelson RJ, Webb NM. Generalizability theory: a primer. Sage Publications; 1991.
[25] Bloch R, Norman G. G_String: a Windows wrapper for urGENOVA. 6.1.1 ed. Hamilton, Canada: The Program for Educational Research and Development (PERD), Faculty of Health Sciences, McMaster University; 2015.
[26] Norman GR, Streiner DL. Biostatistics: the bare essentials. Bc Decker Hamilton; 2008.
[27] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977:159–74.
[28] Hinkle DE, Wiersma W, Jurs SG. Applied statistics for the behavioral sciences. Boston, Mass., [etc.]: Houghton Mifflin; 2003.
[29] Kelley K, Lai K. MBESS. 2015. 3.3.3 ed.
[30] Dunn TJ, Baguley T, Brunsden V. From alpha to omega: a practical solution to the pervasive problem of internal consistency estimation. Br J Psychol 2014;105(3):399–412.
[31] Sijtsma K. On the use, the misuse, and the very limited usefulness of Cronbach's alpha. Psychometrika 2009;74(1):107–20.
[32] Green SB, Yang Y. Commentary on coefficient alpha: a cautionary tale. Psychometrika 2009;74(1):121–35.
[33] Green SB. A coefficient alpha for test-retest data. Psychol Methods 2003;8(1):88.
[34] Revelle W. Hierarchical cluster analysis and the internal structure of tests. Multivariate Behav Res 1979;14(1):57–74.
[35] Hecimovich MD, Styles I, Volet SE. Development and psychometric evaluation of scales to measure professional confidence in manual medicine: a Rasch measurement approach. BMC Res Notes 2014;7(1):338.

[36] Downing SM. Reliability: on the reproducibility of assessment data. Med Educ 2004;38(9):1006–12.
[37] Ziegler M, Poropat A, Mell J. Does the length of a questionnaire matter? J Indiv Differ 2015;35:250–61.
[38] Mazor K, Clauser B, Cohen A, Alper E, Pugnaire M. The dependability of students' ratings of preceptors. Acad Med 1999;74(10):S19–21.
[39] Hoyt WT. Rater bias in psychological research: when is it a problem and what can we do about it? Psychol Methods 2000;5(1):64.
[40] Solomon DJ, Speer AJ, Rosebraugh CJ, DiPette DJ. The reliability of medical student ratings of clinical teaching. Eval Health Prof 1997;20(3):343–52.
[41] Smith CA, Varkey AB, Evans AT, Reilly BM. Evaluating the performance of inpatient attending physicians. J Gen Intern Med 2004;19:766–71.
[42] Beckman TJ, Lee MC, Rohren CH, Pankratz VS. Evaluating an instrument for the peer review of inpatient teaching. Med Teach 2003;25(2):131–5.
[43] Lases SL, Arah OA, Pierik ER, Heineman E, Lombarts MK. Residents' engagement and empathy associated with their perception of faculty's teaching performance. World J Surg 2014;38(11):2753–60.
[44] Brown T, Williams B, Lynch M. Relationship between clinical fieldwork educator performance and health professional students' perceptions of their practice education learning environments. Nurs Health Sci 2013;15(4):510–7.
[45] Mackillop L, Parker-Swift J, Crossley J. Getting the questions right: non-compound questions are more reliable than compound questions on matched multi-source feedback instruments. Med Educ 2011;45(8):843–8.
[46] Bierer SB, Hull AL. Examination of a clinical teaching effectiveness instrument used for summative faculty assessment. Eval Health Prof 2007;30(4):339–61.
[47] van der Leeuw R, Lombarts K, Heineman MJ, Arah O. Systematic evaluation of the teaching qualities of obstetrics and gynecology faculty: reliability and validity of the SETO tools. PloS One 2011;6(5):e19142.
[48] McGrath C, Yeung RWK, Comfort M, McMillan A. Development and evaluation of a questionnaire to evaluate clinical dental teachers (ECDT). Br Dent J 2005;198(1):45–8.
[49] Saarikoski M. Mentor relationship as a tool of professional development of student nurses in clinical practice. Int J Psychiatr Nurs Res 2003;9(1):1014–24.
[50] Ko CY, Escarce JJ, Baker L, Sharp J, Guarino C. Predictors of surgery resident satisfaction with teaching by attendings: a national survey. Ann Surg 2005;241(2):373.
[51] Zinbarg RE, Yovel I, Revelle W, McDonald RP. Estimating generalizability to a latent variable common to all of a scale's indicators: a comparison of estimators for ωh. Appl Psychol Meas 2006;30(2):121–44.
[52] Tavakol M, Dennick R. Making sense of Cronbach's alpha. Int J Med Educ 2011;2:53–5.
[53] Vaughan B, Macfarlane C, Florentine P. Clinical education in the osteopathy program at Victoria University. Int J Osteopath Med 2014;17(3):199–205.