2017-11

Application of Rasch analysis in the development and psychometric evaluation of dental undergraduates preparedness assessment scale

Ali, Kamran

10.1111/eje.12236
European Journal of Dental Education
Wiley

"This is a final author's draft of the paper submitted for publication in European Journal of Dental Education 2016: doi: 10.1111/eje.12236

Title:

APPLICATION OF RASCH ANALYSIS IN THE DEVELOPMENT AND PSYCHOMETRIC EVALUATION OF DENTAL UNDERGRADUATES PREPAREDNESS ASSESSMENT SCALE (DU-PAS)

K. Ali[1], A. Slade[2], E.J. Kay[3], D Zahra[4], A Chatterjee[5], C. Tredwin[6]

[1] Associate Professor and Consultant in Oral Surgery, Peninsula Dental School, Plymouth University,[2] ITM Research Fellow, Primary Care Clinical Sciences, Institute of Applied Health Research, College of Medical and Dental Sciences, University of Birmingham, [3] Foundation Dean, Peninsula Dental School, Plymouth University, [4] Senior Psychometrician Plymouth University Peninsula Schools of Medicine and Dentistry, [5] Lecturer in Technology Enhanced Learning, Plymouth University Peninsula Schools of Medicine, [6]Head of Peninsula Dental School, Plymouth University

Correspondence Author: Kamran Ali

Peninsula Dental School
Portland Square
University of Plymouth
Drake Circus
Plymouth
PL4 8AA
E-mail: kamran.ali@plymouth.ac.uk

ABSTRACT

Aims: The aim of this study was to develop a valid and reliable scale to measure preparedness of new dental graduates.

Methods: The scale development and validation was done using the Rasch measurement model. Following a pilot and pretesting of the scale, a national study was undertaken with undergraduate students from all dental schools as well as foundation dentists in UK.

Results: To examine the internal validity of the scale we conducted a Rasch analysis. External validity of the scale was checked through validation with a range of stakeholders. An excellent fit to the Rasch model provided evidence of internal construct validity. The scale demonstrated invariance, ordered thresholds and lack of differential item functioning. Unidimensionality of the scale was confirmed by independent *t*-tests. The PSI value was 0.877, indicating a good degree of person separation and internal consistency. Test retest reliability of the scale was also established.

Conclusions: The preparedness scale developed in this project reflects innovative research using a systematic approach and employment of modern psychometric methods. The scale can be used for assessment of the preparedness of undergraduate students by dental educators and potential employers as well as by the student for self-assessment.

Key Words: Assessment, Dental, Preparedness, Students

INTRODUCTION

The ultimate goal of undergraduate dental education is to prepare students for a career in dental practice. Preparedness of dental graduates reflects the knowledge, skills and attitudes required to practise dentistry safely, effectively and professionally in a practice environment. Dental educators in Europe have agreed that a graduating dentist "*must have acquired this ability through the achievement of a set of generic and subject specific competences – abilities essential to begin independent, unsupervised dental practice*" (1).

Studies measuring preparedness of dental graduates appear to have focussed primarily on clinical skills of new graduates (2-5). Notwithstanding the importance of clinical skills, they only represent one of several dimensions of preparedness (1, 6). Preparedness is a latent construct as it not directly observable (7). By examining the responses of a participant to a set of items related to the underlying construct, investigators can assign a score that approximates the person's "level" of ability on the latent trait (8).

Item response theory (IRT) measurement models are gaining popularity in the assessment of medical students (9). These models aim to explain the relation between observed score, and an underlying construct, that is, the difference between an item's location and the person's ability (8, 9). The Rasch model tests the observed response patterns against the expectations of the model, a probabilistic form of Guttman scaling (10). Guttman scaling is a deterministic pattern that expects a hierarchical ordering of items based on difficulty so that affirmation of a difficult item will also result in affirmation of an easy item (11). The point at which positive affirmation changes to a negative response affirms the location of the respondent on

the assessment. This approach allows the respondent ability and item ability to be expressed on the same scale of logits. Zero on the scale represents the centre of the item difficulty range while the respondent can be on either side of zero allowing measurement of respondent ability and item difficulty independently from each other.

Rasch analysis allows evaluation of several key psychometric properties of a scale (12-14). These include: the *overall fit statistics* of the scale to the model to evaluate *invariance* or stability properties of the scale; *response thresholds* for each item on a scale; *differential item functioning* to assess bias in sub-groups; r*esponse dependency*; and *unidimensionality* of a scale.

The aim of this study was to develop a valid and reliable scale to measure preparedness of new dental graduates using Rasch analysis.

METHODS

Ethical approval for this study was obtained from the institution research ethics committee. The construct of preparedness was operationalised by identifying a pool of potential latent construct indicators. A total of 80 areas encompassing the knowledge, clinical skills and behavioural attributes expected from new graduates were identified from triangulation of qualitative research and existing literature. Pretesting of items was carried out to ensure that scale items were comprehensible, unambiguous and scoring categories could be interpreted with ease. The content and face validity of scale items was established with a range of stakeholders during pre-testing. Dental students (N=10) and dental experts (N=12) participated in open and closed ended questions regarding the clarity, representativeness, relevance, and distribution (difficulty) of the scale. This was followed by cognitive interviews based on verbal probing. Following pretesting, a number of items were re-worded and two redundant items were deleted. Finally, 78 items were identified for potential inclusion in the pilot phase. The 78-item inventory was divided into three versions with 36 items designated as *core linking items* and 14 additional items in each version leading to a total of 50 items per scale.

An online pilot study was undertaken to confirm the feasibility of the study. BDS students in Year three and Year two from a dental school in the South West of England (N=64) participated in the pilot. The web-addresses for the three versions of the scale were embedded in a single web-link sent as an e-mail invitation to participants. Clicking on the link took participants to the website and the web-link automatically rotated between the three different versions, allowing all three versions to be attempted equally in succession. However, each participant was only required to complete one version of the scale.

Data analysis of the pilot study highlighted the strengths and weaknesses of the scale and allowed informed decisions to improve the scale structure. Items with low discrimination and extreme fit statistics were deleted along with revisions of the outcome space. The post pilot scale inventory comprised 62 items distributed in three versions and included 35 core items and nine additional items in each version.

Following the pilot study, a national study was undertaken involving undergraduate dental students and new graduates undertaking dental foundation training. Recruitment of dental undergraduates was done through the Dental Schools Council (DSC) while Foundation Dentists were contacted through the Committee of Postgraduate Dental Deans and Directors (COPDEND).

Finally, test retest reliability of the scale was also assessed. Participants were invited to complete the questionnaire again within two weeks of the initial assessment.

Data Analysis

Rasch analysis was conducted using Rasch unidimensional measurement model software, RUMM2030 *(*Perth, Western Australia: RUMM Laboratory Pty Ltd, 2010). The Unrestricted, Partial Credit Model was used.

RESULTS

Participants

A sample of 457 participants accessed the web-link and of these 392 participants provided complete responses (85.77%). The gender distribution of the participants included 236 females (60.20%) and 156 males (39.80%). The majority of participants (88.30%) were in the 20-29-year age group. With regards to their current professional status, 196 participants were undergraduate dental students (Final year and pre-final year) while 191 were foundation dentists.

Rasch Analysis

Initial Analysis

Data from the three versions of the scale was combined into a single data set for Rasch analysis. The initial results identified a number of issues with the psychometric properties of the scale including lack of invariance, items with extreme fit residuals and disordered thresholds. The overall test fit (item-trait interaction) was significant (x^2=490.43; d.f.=305; p<0.001) indicating the hierarchical ordering of the items across the trait showed lack of invariance, and suggesting some degree of misfit to the Rasch model.

This could be caused by misfit to model expectations of items or respondents, or both. This warranted further investigations to identify the source(s) of misfit. Firstly, the threshold map identified 11 items with disordered thresholds. Item fit identified four items were with fit residuals greater than ±2.5 and significant chi squares.

Several steps were taken to address these issues through iteration and repetition of analysis. Two key approaches which were utilised to address the problems identified

included rescoring of response categories and deletion of items. The main reasons for deletion of items were correction of disordered thresholds persisting after rescoring of response categories, extreme fit residuals outside the ±2.5 range, item dependency and low discrimination. In total 12 items were deleted. The revised version of the scale consisted of 50 items with 3 response categories across the entire length of the scale (scored as 0, 1 and 2) and yielding a maximum raw score of 100 (Appendix 1).

Revised Scale

Following revisions, summary statistics were computed for the revised version of the scale. Overall Rasch chi-square fit statistics for the refined preparedness scale was not significant (x^2=272.55; d.f.=250; *p*= 0.156), indicating an adequate fit to the Rasch model. The PSI value was 0.877, indicating a good degree of person separation and internal consistency. A PSI at this level indicated that the scale was able to discriminate between three or more ability groups of respondents.

The *persons fit statistics* (mean person location) showed a mean of 3.905; SD ±1.435 indicating that, in general, the response group was of a higher ability level than the item difficulty of the preparedness scale (Mean item location=0.000; SD ±1.869). The person-item threshold distribution is depicted in Figure 1. The graph shows the distributions of persons and item thresholds for the finalised 50-item preparedness scale. It supports the observations in the preceding paragraph, that is, the person ability was higher than the mean item difficulty. The SD of summary *fit residual statistics* was within an acceptable range for items (1.063) as well as persons (0.811) and indicated adequate fit to the model.

Logit scores (interval-level) were converted back into the original (raw) score range of the scale (0-100) for ease of interpretation. A raw score of zero was equal to -8.645 logits while a score of 100 translated into 7.978 logits. The mean person location of 3.905 was equivalent to a raw score of 75.

The location differences between the person-factor groups were investigated. The mean location for female students was 3.810; SD ±1.43 and was 4.048; SD ±1.44 for male students. However, ANOVA statistics showed no significant effects for gender [F (1, 390) =2.59, *p*= 0.108]. Next, the effect of age on person-item threshold distribution was investigated. ANOVA statistics showed there was no significant effect of age on person location [F (3, 388) = 1.462, *p*=0.224]. A comparison of scores based on the professional status of the respondents showed that foundation dentists were more likely to have higher logit values (mean person location=4.204) when compared to final year students (mean person location=4.044 and pre-final year students (mean person location=2.339). ANOVA statistics confirmed these differences were significant [F (3, 388) = 28.116, *p*=0.000].

Item fit statistics of the revised scale showed that all items, except one showed adequate fit to the model expectations with fit residuals ± 2.5. One item displayed a slightly higher fit residual of 2.60 suggesting a low level of discrimination. Examination of the category probability curves showed that a small number of respondents in the extreme age groups were skewing the responses. Deletion of respondents aged less than 20 years (N=4) and over 40 years (N=4) resulted in the fit residual for item B060 dropping to an acceptable 2.45.

The revised *item threshold map* showed ordered thresholds for all 50 items. The threshold map in item location order is depicted in Figure 2. Respondents with higher

ability endorsed difficult items, while individuals with low ability consistently endorsed less difficult items. Inspection of the threshold map showed that the threshold distances vary across items, supporting the use of the partial credit model for the analysis of this scale.

Differential item functioning (DIF) of the revised scale was assessed to identify any bias for person factors including gender and age.

Firstly, DIF was investigated for gender and statistics were computed for both *Uniform* DIF as well as *non uniform* DIF. The results confirmed the scale to be free of gender bias as no uniform or non-uniform DIF by gender was identified. DIF was investigated for age and the results confirmed the absence of uniform DIF. However, one item showed presence of non-uniform DIF ($p<0.05$). Investigations identified the persons in age-group less than 20 years (N=4) as the source of DIF. Given the small number of respondents in this group, it was deemed appropriate to delete this group of respondents. Following this, non-uniform DIF was eliminated from all items ($p=0.999$).

Local dependency of items was investigated by examining the residual correlations (>0.2). Several items in the revised scale showed local dependency and included the following:

- Items related to radiography and radiology skills (A003, A004, A005, and A006),
- Items related to endodontics (A024 and A025),
- Items related to removable prosthesis (A028 and A029)
- Items related to safety in practice (B036 and B040).
- Items related to communication skills (B045, B046, B047, B048 and B049).

The rationale for retaining items in the revised scale is explained in the discussion section.

Unidimensionality of the scale was computed using the residuals of the principal component analysis (PCA). Person estimates, based on items displaying high positive and high negative loadings on the first principal component (PC1) of the residuals were tested for significant differences. A series of independent *t*-tests comparing subsets identified from PCA analysis of residuals showed only 2.36% tests to be significant, confirming unidimensionality of the scale.

Test-re test Reliability

In total 76 participants responded to retest invitations. The overall raw scores on test occasion 1 and 2 for each participant were analysed in SPSS. However, 12 participants with missing data points were deleted, leaving 64 participants in the data analysis. The results shown in Table 2 indicated an acceptable level of test-retest reliability (Pearson's r>0.783), significant at the 0.01 level (2-tailed) indicating acceptable correlations between scores on test occasion 1 and 2.

Test retest reliability statistics were computed based on logits (interval-level scores) for the 50 items in the finalised scale. The logit scores on test occasion 1 and 2 were transformed across the full range of the original raw score for the preparedness scale (0-100). The results showed excellent internal consistency (Cronbach's alpha=0.916). A significant correlation between scores on test occasion 1, and 2 at the 0.01 level (2-tailed) was observed (Pearson's r = 0.847). The test-retest reliability statistics indicated good reproducibility of the scale with low level of random measurement error.

DISCUSSION

Scale development is a challenging task and entails careful planning and attention to detail (15, 16). The use of appropriate models such as the Rasch measurement model offers a more precise measurement of latent traits than can be achieved with other approaches. The model operationalises the formal axioms which underpin measurement allowing testing of items in a scale for fit to the Rasch model (17). The mathematical model underlying Rasch analysis is a special case of IRT and provides a useful approach for assessing the psychometric properties of categorical data as a function of trade-off between respondent's abilities and item difficulty. This approach helps overcome the limitations of traditional measurements based on Classical Test Theory (18). IRT models such as Rasch analysis have not been used extensively in healthcare education and this may be primarily attributed to their relatively complex nature (19).

The scale developed in this study satisfies the expectations of the Rasch model and provides evidence of internal construct validity as shown by an excellent fit to the Rasch model, adequate person separation index, ordered thresholds, unidimensionality, and lack of DIF. Test retest reliability of the scale was established using contemporary guidelines (15, 16) yielding good reproducibility.

The differences in the mean person ability and the mean item difficulty scores may suggest poor targeting. However, this can be explained, given the learning outcomes and standards of practice, dental schools need to sign off all dental students to be competent in a range of core skills and attributes prior to graduation to ensure patient safety and protection of the public. The mean person location of 3.905 logits on this scale translates into a raw score of 75 out of 100 (maximum score) which is not high in the context of graduating dentists. Some of the items such as

taking a medical history appeared to be at the floor of the scale suggesting redundancy. Nevertheless, it is vital to assess core skills. Retaining these items in the scale means it can be used to in the future to monitor progress and the scale is then robust enough to pick up skills acquired early on in a student's studies.

In addition, endorsement of the scale items was based on self-assessment possibly inflating the mean person ability. Evidence shows poor correlations between perceived self-confidence and observed competence (20, 21). Perhaps, the next step may be to compare the scores of self-assessment with the assessment by dental educators, which may provide a more realistic measurement, enabling tutor feedback and identifying areas of need. Finally, it has been demonstrated that a sample size of 243 is adequate to give a degree of precision of ± 0.5 logits, at 99% confidence, even when the targeting of the scale is not good (22). The sample size in this study was sufficient for a precise measurement despite the targeting being less than desirable.

Several items in the scale demonstrated local dependency. However, local dependency should be interpreted with caution in order to avoid any compromise in the content validity of the preparedness assessment scale. If each item is aimed at assessing a somewhat different aspect of the trait, retaining locally dependent items can provide greater information about the ability of the students and maintain the content validity (23). Therefore, items assessing related but distinct skills and attributes were retained.

The preparedness scale offers several potential applications, including assessment of undergraduate dental students by dental educators and potential employers as well as for self assessment. These assessments can be done periodically throughout

the undergraduate programmes to map the progress of dental students longitudinally with a final assessment at the time of graduation. It can also be used to make useful comparisons of the ability level of a student as measured by dental educators and student's self-assessment. These can be utilised for two-way feedback, identifying strengths and weaknesses at each stage of the BDS programme and informing appropriate remediation and consolidation. Dental educators can also use it for comparisons at an inter-institutional level and the data thus generated can be used to inform teaching and learning strategies at dental schools.

In regards to the limitations, the scale is primarily aimed at measuring preparedness of new graduates in the UK. If it is to be used in other countries, they would need to ascertain its external validity and evaluate its relevance, adequacy and representativeness in the context of the learning outcomes of undergraduate dental programmes. The practice of dentistry and the role of dentists is constantly evolving (24, 25). The dental profession is sensitive to scientific and educational developments as well as socio-economic influences and they may impact on the perceptions of stakeholders over time. Therefore, the content validity of the scale will need to be monitored periodically to ensure that it remains appropriate for use with the introduction of new regulations and advancements in clinical care.

CONCLUSION

A methodical approach was used to develop a valid and reliable scale to measure preparedness of new dental graduates using modern psychometric methods. The psychometric properties of the preparedness scale conformed to the Rasch model providing evidence for its unidimensionality and ability to provide an interval-level measurement. The preparedness scale offers promise for its use in the assessment of dental students and new graduates.

Acknowledgements

The authors would like to thank the Association for Dental Education in Europe (ADEE) for supporting this study through the ADEE 40th Anniversary Scholarship; Dental Schools Council, UK and the Committee of Postgraduate Dental Deans and Directors (COPDEND) for facilitating a national study. We are also thankful to all the participants for their contribution to this research.

Disclosure

None of the authors have any conflict of interest to declare.

Figure 1 Person-item threshold distribution of revised scale

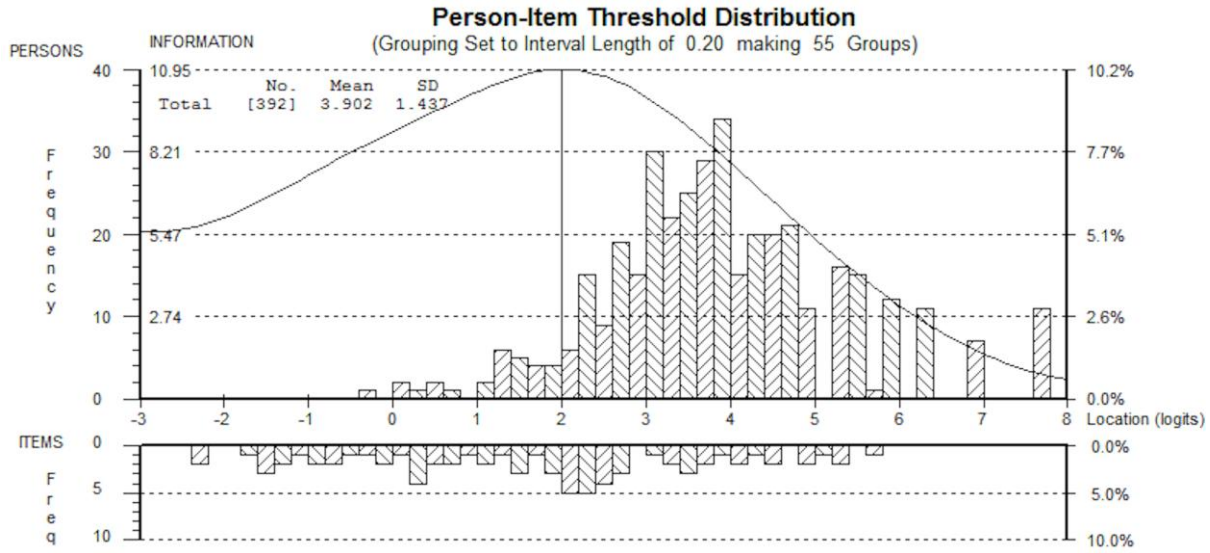

Figure 2 Item threshold map of revised scale

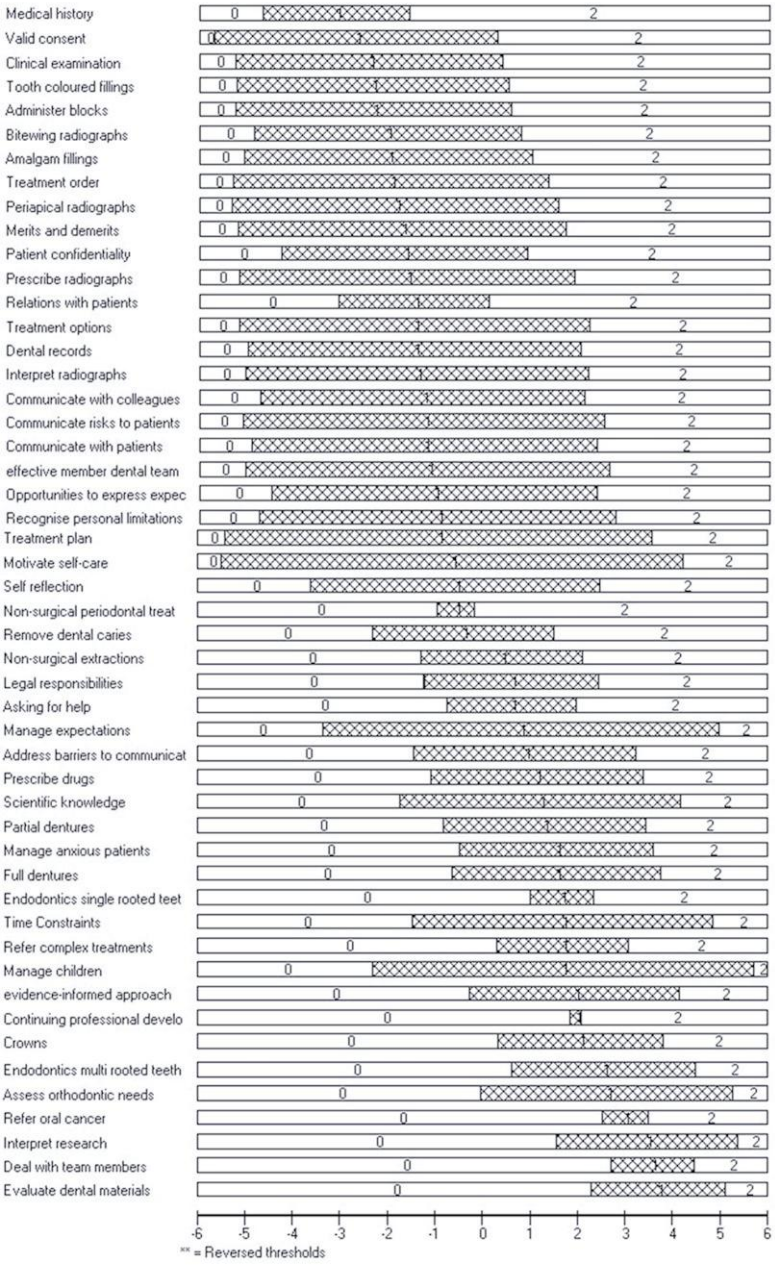

Table 1: Test-rest reliability statistics

	Version A	Version B	Version C	Core 35
N	19	20	25	64
Pearson's *r*	0.925	0.783	0.896	0.898
p	<0.001	<0.001	<0.001	<0.001
Partial* *r*	0.913	0.833	0.898	0.904
P	<0.001	<0.001	<0.001	<0.001

*Controlling for gender, age group, and status group

REFERENCES

1. Cowpe J, Plasschaert A, Harzer W, Vinkka-Puhakka H, Walmsley AD. Profile and competences for the graduating European dentist–update 2009. European Journal of Dental Education. 2010;14(4):193-202.
2. Gilmour A, Jones R, Bullock A. Dental foundation trainers' expectations of a dental graduate: final report. Wales Deanery/Cardiff University. 2012.
3. Greenwood LF, Townsend GC, Wetherell JD, Mullins GA. Self-perceived competency at graduation: a comparison of dental graduates from the Adelaide PBL curriculum and the Toronto traditional curriculum. European Journal of Dental Education. 1999;3(4):153-8.
4. Yiu C, McGrath C, Bridges S, Corbet E, Botelho M, Dyson J, et al. Self-perceived preparedness for dental practice amongst graduates of The University of Hong Kong's integrated PBL dental curriculum. European Journal of Dental Education. 2012;16(1): e96-e105.
5. Manakil J, George R. Self-perceived work preparedness of the graduating dental students. European Journal of Dental Education. 2013;17(2):101-5.
6. General Dental Council. Preparing for practice. Dental team learning outcomes for registration London: GDC. 2011.
7. Streiner DL, Norman GR, Cairney J. Health measurement scales: a practical guide to their development and use. Oxford University Press, USA; 2014.
8. Sharkness J, DeAngelo L. Measuring student involvement: A comparison of classical test theory and item response theory in the construction of scales from student surveys. Research in Higher Education. 2011; 52(5):480-507.
9. Newby VA, Conner GR, Grant CP, Bunderson CV. The Rasch model and additive conjoint measurement. J Appl Meas. 2009; 10(4):348-54.
10. Rasch, G. (1960/1980) *Probabilistic models for some intelligence and attainment tests.* Danish Institute for Educational Research, T.U.o.C.P. Copenhagen, Chicago: Danish Institute for Educational Research, The University of Chicago Press.
11. Tennant A, McKenna SP, Hagell P. Application of Rasch analysis in the development and application of quality of life instruments. Value in Health. 2004;7(s1): S22-S26.
12. Hagquist C, Bruce M, Gustavsson JP. Using the Rasch model in nursing research: an introduction and illustrative example. International Journal of Nursing Studies. 2009;46(3):380-93.
13. Tennant A, Conaghan PG. The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? Arthritis Rheum. 2007 Dec 15;57(8):1358-62. Review.
14. Shea TL, Tennant A, Pallant JF. Rasch model analysis of the Depression, Anxiety and Stress Scales (DASS). BMC psychiatry. 2009;9(1):21.
15. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Quality of Life Research. 2010;19(4):539-49.
16. Artino Jr AR, La Rochelle JS, Dezee KJ, Gehlbach H. Developing questionnaires for educational research: AMEE Guide No. 87. Medical teacher. 2014; 36(6):463-74.

17. Wolfe E, Smith Jr E. Instrument development tools and activities for measure validation using Rasch models: part I-instrument development tools. J Appl Meas. 2006; 8(1):97-123.
18. Hagquist C, Bruce M, Gustavsson JP. Using the Rasch model in nursing research: an introduction and illustrative example. Int J Nurs Stud. 2009 Mar;46(3):380-93.
19. Tavakol M, Dennick R. Psychometric evaluation of a knowledge based examination using Rasch analysis: An illustrative guide: AMEE Guide No. 72. Medical teacher. 2013; 35(1): e838-e48.
20. Colthart I, Bagnall G, Evans A, Allbutt H, Haig A, Illing J, et al. The effectiveness of self-assessment on the identification of learner needs, learner activity, and impact on clinical practice: BEME Guide no. 10. Medical teacher. 2008;30(2):124-45.
21. Barnsley L, Lyon PM, Ralston SJ, Hibbert EJ, Cunningham I, Gordon FC, et al. Clinical skills in junior medical officers: a comparison of self-reported confidence and observed competence. Medical education. 2004;38(4):358-67.
22. Linacre J. Sample Size and Item Calibration Stability: Rasch Measurement Transactions 1994.7: 4 p. 328; Australia.
23. Andrich D, Kreiner S. Quantifying response dependence between two dichotomous items using the Rasch model. Applied Psychological Measurement. 2010;34(3):181-92.
24. Kanaparthy R, Kanaparthy A. The changing face of dentistry: nanotechnology. International journal of nanomedicine. 2011; 6:2799.
25. Fincham AG, Shuler CF. The changing face of dental education: the impact of PBL. Journal of Dental Education. 2001;65(5):406-21.