DEVELOPMENT AND VALIDATION OF A SCALE EVALUATING TEACHING LEVEL AT A MEDICAL UNIVERSITY (ETLTS-MU)：COMBINATION OF A CLASSICAL TEST THEORY AND GENERALIZABILITY

Qiong Meng, Ying Chen, Liping He, Jianzhong Yin, Jianyun Yu, Yanchun Sun
Kunming Medical University, China
Ping Chen
Yunnan Health Education Institute, China

Abstract

Objective: *To develop the scale evaluating the level of teaching at a medical university (ETLTS-MU) and validate it by both Classical Test Theory (CTT) and Generalizability Theory (GT).*

Methods: *The ETLTS-MU was developed based on programmed decision procedures with three rounds of discussions, in-depth interviews and quantitative statistical procedures. The psychometric properties of the scale were evaluated with respect to validity and reliability employing correlation analysis, factor analysis and also G-study and D-study of GT analysis.*

Results: *The total of 485 students were interviewed using ETLTS-MU and data from 477 effective questionnaires were analyzed. Correlation and factor analyses confirmed good construct validity. The Cronbach's α coefficient of the overall scale was 0.967 and the Cronbach's α coefficient of each domain was larger than 0.7. In GT, generalizability coefficients and indexes of dependability confirmed the reliability of the scale further with more exact variance components.*

Conclusions: *The ETLTS-MU has good validity, reliability and some highlights, and can be used as a tool for evaluating the level of teaching at a medical university. However, in order to obtain better reliability, the numbers of items for teaching organization dimension should be increased or the quality of the items of this dimension should be improved.*

Keywords
Teaching level, Scale, Reliability, Validity, Generalizability theory

Background

The level of teaching at universities and colleges implies the teaching ability and the level of teachers involved in the the teaching process at universities and colleges. The evaluation from students can better reflect the teaching level because students take part in the teaching and learning process all thorugh the study period and are the direct beneficiaries (Kaiyong Tang, 2003). The earliest study on evaluation of the teaching level at universities and colleges was a research into the teaching quality evaluation carried out at Beijing Normal University in 1984.The evaluation of the teaching level had been carried out at most universities and colleges in China until 1995 (Chunlin Li, 2005). Developing a practical, reasonable and efficient instrument with medical characteristic is the base of evaluation of the teaching level at medical universities and colleges. Bearing this in mind, we attempted to describe the developmental process of the scale evaluating the teaching level viewed by students of a medical university (ETLTS-MU) and to analyze the validity and reliability of this scale by applying classical test theory (CTT) and further analyze the reliability of this theory by applying generalizability theory (GT).

Methods

Establishment of the scale (ETLTS-MU)

A nominal group consisting of 10 individuals (5 teachers and 5 students) and a focus group with 5 experts (3 teachers and 2 teaching management personnel) were organized to present the conceptual framework and select items by using the programmed decision method. A definite framework of the ETLTS-MU scale was put forward by the focus group after discussion and the scale should be divided into 5 dimensions, including teaching organization, teaching contents, teaching methods, teaching attitude, and teaching effect. The nominal group proposed items for each dimension to form the item pool based on the framework and fully considering the medical characteristic. The item selection was based not only on qualitative analysis such as nominal group, focus group discussions and in-depth interviews, but also on four quantitative statistical procedures. The main steps of developing the ETLTS-MU were summarized as a schematic diagram below:

Organizing two groups
↓ focus group discussions
A definite framework
↓ nominal group proposed items
Item pool (50 items)
↓ In-depth interview, focus and nominal group discussions, four quantitative statistical procedures
Screened refining items (30 items)
↓ analysis, focus group discussions
Final scale (5 dimensions, 24 items)

The final ETLTS-MU included 24 items which are classified into 5 dimensions with teaching organization dimension including 3 items (coded TO1-TO3), teaching contents dimension including 7 items (coded PS1-PS7) and teaching methods dimension including 5 items (coded TM1-TM5), teaching attitude dimension including 5 items (coded TA1- TA5), and teaching effect dimension including 4 items (coded TE1-TE4).

Validation of the ETLTS-MU

Data collection and scoring

The formal ETLTS-MU described above was used for students who were taught by 10 teachers from the School of Public Health at Kunming Medical University in a field survey in order to study its psychometric properties (validity and reliability). Each student was asked to answer every item of the scale after the participating investigators explained the process and the scale. The resoponses were checked immediately each time by the investigators to make sure that the questionnaires are complete. If missing values were found, a questionnaire would be returned to a student to fill in the missing item.

Based on the data collected, the raw scores of items, dimensions and overall scale were calculated. Each item of ETLTS-MU was rated in a five-level Likert scoring system, namely, *not at all, a little bit, somewhat, quite a bit, and very much*. Each dimension score was obtained by adding together the within-dimension item scores. The overall scale score was the sum of the five dimension scores. For comparison purposes, all domain scores were linearly converted to a 0–100 scale using the formula: SS = (RS-Min) × 100/R, where SS, RS, Min and R represented the standardized score, raw score, minimum score, and a range of scores, respectively.

Statistical analysis for psychometrics

Validity is the degree to which the instrument measures what it is supposed to measure (Liabing Qi et al, 2003). Construct validity was evaluated by Pearson's correlation coefficient r (item-dimension correlations) as well as by factor analysis with varimax rotation (Chaojie Liu,1997) . Reliability is the degree to which an instrument is free from random errors, with being evaluated by measuring internal consistency reliability in our research. The internal consistency, which refers to the homogeneity of the items of the scale, was assessed by Cronbach's alpha coefficient for each dimension. A high internal consistency suggests that the scale is measuring a single construct (Can Li, 2008).

Generalizability theory analysis

Besides CTT analysis above, we also applied GT to investigate the score dependability of the ETLTS-MU scale. GT, one kind of the modern measuring theory, can decompose and control all kinds of errors generated from the measurement by applying variance analysis techniques and the thought of experimental design (Zhiming Yang and Lei Zhang, 2003). GT contains two stages: Generalizability Study (G-study) and Decision Study (D-study). G-study serves as a "pilot" study that decomposes the variance and covariance components related with various error sources in help to confirm the relationship between measurement goal and measurement facets based on the data collected by using analysis of variance (ANOVA). In D-study, the information from the G-study (Shavelson RJ and Webb NM, 1991) is used for planning of “optimal” measurement protocol so that the best possible reliability can be achieved while balancing other factors by calculated two reliability coefficients: generalizability coefficient (*G*) and index of dependability (*Φ*).GT has been presented as a way to refine the designs of measurement procedures in an attempt to yield reliable data (Winterstein BP et al ,2010) (Stora B et al, 2013)(Crits-Christoph P et al, 2011) (Heitman RJ et al,2009) (Cella DF et al, 1993). We defined the teaching level as the target of measurement and items as one facet of a measurement error. Given every student is asked to reply to all items, the design is single-facet crossed design, namely $p \times i$ design.

Results

The total of 485 questionnaires were sent out and 477 questionnaires were effective. All the students completed the ETLTS-MU scale in 5 minutes. Data from effective questionnaires were used to analyze the validity and the reliability of this scale by using CTT. For simplifying the design of GT, 186 effective questionnaires from the same class in which one teacher taught were used to further analyze the reliability by using GT.

Construct validity

Correlation analyses showed that there were strong associations between the items and their own dimensions (all correlation coefficients are higher than 0.5), but weak relationship between items and other dimensions (listed in Table 1). For example, correlation coefficients between TOD and items of TO1-TO3 (in bold) were higher than those between TOD and other items.

TOD, teaching organization dimension; TCD, teaching contents dimension; TMD, teaching methods dimension; TAD teaching attitude dimension; TED, teaching effect dimension

There were 5 principal components (initial eigenvalues>1) abstracted from 24 items by factor analysis, accounting for 75.68% of the cumulative variance. By using the varimax rotation method, it can be seen that the 5 principal components reflected different facets under five dimensions. Specifically, the first principal component mainly represented the teaching methods dimension and the teaching effect dimension with higher factor loadings; the second principal component, the third principal component and the fourth principal component mainly reflected the teaching attitude dimension, teaching contents dimension and teaching organization dimension with higher factor loadings respectively.

Table 1 Item- dimension correlation coefficients for the ETLTS-MU scale

Item	TOD	TCD	TMD	TAD	TED
TO1	**0.594**	0.535	0.458	0.447	0.411
TO2	**0.871**	0.616	0.638	0.514	0.589
TO3	**0.913**	0.646	0.695	0.533	0.629
TC1	0.473	**0.633**	0.470	0.497	0.475
TC2	0.555	**0.764**	0.603	0.598	0.576
TC3	0.501	**0.716**	0.533	0.541	0.535
TC4	0.496	**0.727**	0.559	0.565	0.530
TC5	0.540	**0.782**	0.584	0.558	0.555
TC6	0.582	**0.745**	0.590	0.897	0.538
TC7	0.635	**0.832**	0.698	0.559	0.663
TM1	0.630	0.643	**0.858**	0.551	0.701
TM2	0.673	0.676	**0.895**	0.596	0.733
TM3	0.634	0.687	**0.855**	0.626	0.724
TM4	0.563	0.625	**0.716**	0.617	0.639
TM5	0.619	0.650	**0.848**	0.573	0.729
TA1	0.440	0.582	0.523	**0.721**	0.556
TA2	0.452	0.577	0.512	**0.759**	0.543
TA3	0.467	0.592	0.559	**0.737**	0.549
TA4	0.479	0.573	0.536	**0.712**	0.545
TA5	0.508	0.512	0.586	**0.795**	0.575
TE1	0.578	0.630	0.727	0.621	**0.897**
TE2	0.596	0.621	0.725	0.578	**0.883**
TE3	0.610	0.663	0.743	0.617	**0.870**
TE4	0.536	0.614	0.626	0.581	**0.741**

Reliability from CTT

As can be seen in Table 2, the Cronbach's α coefficient for each dimension was higher than 0.75 and the Cronbach's coefficient for the overall scale was 0.967.

Table 2 Reliability statistics based on CTT for the ETLTS-MU scale

Dimension	Number of items	Dimension Score (Mean ± SD)	internal consistency coefficient α
Teaching organization dimension(TOD)	3	86.19 ± 15.43	0.757
Teaching contents dimension(TCD)	7	93.66 ± 10.14	0.926
Teaching methods dimension(TMD)	5	86.38 ± 15.35	0.916
Teaching attitude dimension(TAD)	5	93.77 ± 11.13	0.882
Teaching effect dimension(TED)	4	88.90 ± 14.37	0.907
Overall Scale dimension(TSD)	24	90.44 ± 11.21	0.967

Reliability from GT

G-study results were provided in Table 3 based on the current design, in which 186 students filled in ETLTS-MU scale with 24 items. For TOD, the variances accounted for 50% by person and 42% by person-by-item interactions, only a small source of variation (8%) was due to an item. Similarly, the largest source of variation was due to a person in other dimensions, while the smallest source of variation was due to an item.

Table 3 Estimated variance components and percentage of variance for p × i design in G-study for five dimensions of ETLTS-MU

Dimension	*p*(person)		*i*(item)		*p*×*i*(person×item)	
	Variance component	Percent (%)	Variance component	Percent (%)	Variance component	Percent (%)
TOD	0.261	50	0.042	8	0.219	42
TCD	0.132	55	0.008	3	0.099	42
TMD	0.300	63	0.008	2	0.170	35
TAD	0.141	56	0.002	1	0.107	43
TED	0.247	61	0.017	4	0.140	35

p: person effect, *i*: item effect, *p* × *i*: person-by-item interaction effect.

Several multivariate D studies were performed to estimate G and Φ for the current design and alternative designs with varied numbers of items for five dimensions of ETLTS-MU, with results presenting in Table 4. It showed *G* and *Φ* coefficients for four of five domains both were larger than 0.8 except for TOD based on the original test length (in bold). These two reliability coefficients for TOD were larger than 0.70 but smaller than 0.80. In addition, Table 4 showed G and Φ coefficients were increased with the increasing number of items for each dimension.

Table 4 G-coefficients and Φ-coefficients for different numbers of items for *p* × *I* design in D-study for five dimensions of ETLTS-MU

Dimension	Number of items	$\sigma^2 (P)$	$\sigma^2 (I)$	$\sigma^2 (PI)$	$\sigma^2 (\delta)$	$\sigma^2 (\Delta)$	*G*	*Φ*
TOD	2	0.261	0.021	0.110	0.110	0.131	0.704	0.666
	3	**0.261**	**0.014**	**0.073**	**0.073**	**0.087**	**0.781**	**0.750**
	4	0.261	0.011	0.055	0.055	0.065	0.827	0.800
	5	0.261	0.008	0.044	0.044	0.052	0.856	0.833
TCD	2	0.132	0.004	0.050	0.050	0.053	0.728	0.713
	3	0.132	0.003	0.033	0.033	0.036	0.800	0.788
	4	0.132	0.002	0.025	0.025	0.027	0.842	0.832
	7	**0.132**	**0.001**	**0.014**	**0.014**	**0.015**	**0.903**	**0.897**
TMD	2	0.300	0.004	0.085	0.085	0.089	0.779	0.771
	3	0.300	0.003	0.057	0.057	0.059	0.841	0.835
	4	0.300	0.002	0.043	0.043	0.045	0.876	0.871
	5	**0.300**	**0.002**	**0.034**	**0.034**	**0.036**	**0.898**	**0.894**
TAD	3	0.141	0.001	0.036	0.036	0.036	0.798	0.795
	4	0.141	0.000	0.027	0.027	0.027	0.840	0.838
	5	**0.141**	**0.000**	**0.021**	**0.021**	**0.022**	**0.868**	**0.866**
TED	2	0.247	0.008	0.070	0.070	0.078	0.779	0.759
	3	0.247	0.006	0.047	0.047	0.052	0.841	0.825
	4	**0.247**	**0.004**	**0.035**	**0.035**	**0.039**	**0.876**	**0.863**
	5	0.247	0.003	0.028	0.028	0.031	0.898	0.887

$\sigma^2(\delta)$ is the variance components of relative error;$\sigma^2(\Delta)$ is the variance components of absolute error;$\sigma^2(PI)$ is the variance components of error when estimating the universe score by using sample mean; *G* is the Generalizability coefficient; *Φ* is the index of dependability.

Discussions

Advantages of the ETLTS-MU scale

All items of the evaluation scale were put forward by the teachers and students together and the final version was determined by three rounds of discussions and four quantitative statistical procedures. The structure of the scale is clear and the contents of the scale is quite common and easy to understand. It is important that for this study the students did not have to assign a specific score to teachers unlike the previous evaluations. It was easier to provide fair answers and the process was less time consuming because the response option of each item was rated by a five-level scoring system. Therefore, the scale has good feasibility.

Psychometrics of the ETLTS-MU scale

By the programmed decision procedures, we developed the ETLTS-MU by using the focus group discussion, in-depth interview and pre-testing in order to effectively reduce the number of items in the final version to 24 from an original 50 item pool, ensuring good content validity and sound conceptual structure. It is well recognized that internal consistency (α) should be at least 0.70 and reliability (r) should be above 0.80 in a test–retest situation (Terwee CB et al, 2003). Thus, our results in Table 3 showed that this instrument has good internal consistency reliability, for all Cronbach's α coefficients were higher than 0.70. Correlation analyses showed strong correlation between the items and their own dimension but weak correlation between items and other dimensions. Factor analysis revealed that the components extracted from the data basically coincide with the theoretical construct of the instrument. These results confirmed the good construct validity.

Analysis of generalizability theory

The *G* and *Φ* coefficients are the two important reliability coefficients used to depict the reliability of "relative decision" and "absolute decision" in GT. Which coefficients will be selected depending on the researchers' interests? If one's interest lies in ranking people (relative decision), then the G informs about how dependable a score is. If one's interest lies in the absolute standings to a criterion (absolute decision), *Φ* reflects the score dependability. The major objective of this study was to develop the evaluation scale, ETLTS-MU, to evaluate and rank the teaching level, so G should be selected.

Some researchers (Yifang Wu and Hueying Tzou, 2015; Winterstein B P et al, 2010) suggested that the reliability of an instrument is generally good when the reliability coefficients (*G* and *Φ*) were above 0.80 in GT. For the teaching organization dimension, G was 0.781 based on the current design, which was a little below the good level of 0.80. It will be better to increase the numbers of items of TOD from 3 to 4 in order to reach a good dependability. For other dimensions, G were all greater than 0.80 based on the current design. It can be considered that current items are reasonable for these dimensions.

This research showed that both *G* and *Φ* were increased with the increasing of the number of items. However, increasing the number of items might not be realistic in practice because it was possible that the reliability conversely was reduced with too many items and intensive consumption of time. Hence, the number of items of some dimensions can be decreased under the premise of keeping good reliability (G was above 0.80). The following suggestions are provided: the number of items in teaching organization dimension can be increased from 3 to 4; the number of items in teaching contents dimension, teaching methods dimension and teaching effect dimension can be reduced to 3; the number of items in teaching attitude dimension can be reduced to 4; the total number of items for overall scale will be 17.

Conclusions

To sum up, the evaluation scale, ETLTS-MU, has good reliability, validity and feasibility, and can be used as the instrument evaluating the level of teaching at a medical university. However, for obtaining better reliability, the numbers of items for teaching organization dimension should be increased or the quality of the items of this dimension should be improved.

Grants: The study is supported by the graduate course reform program in the emphasis breaking test unit project of the Ministry of Education of China

References

1. *Kaiyong Tang. (2003). Establishing of evaluation system of the teaching level of teachers. Machine Vocational Education, vol.5:p.14～15.*
2. *Chunlin Li. (2005). Analyzes on practical development and theoretical discussion of the teaching level of teachers in university or college in China. Modern education science,vol.4:p.79～81.*
3. *Liabing Qi，Yihui Zhang，Youzeng Zhen. Analysis on the Reliability and Validity of Questionnaire. Contemporary education science2003,2:53～54.*
4. *Chaojie Liu. 1997. The Assessment of the Reliability and Validity of Questionnaire. Chinese Journal of Prevention and Control of Chronic diseases, vol.5:p.174～177.*
5. *Can Li，Lin Xin. 2008. Research on the evaluation methodology about the Reliability and Validity of Questionnaire. Chinese Journal of Health Statistics, vol.25:p.541～544.*
6. *Yang Zhiming, Zhang lei. 2003. Generalizability theory and its applications. Beijing :Educational science Publishing House, pp.24.*
7. *Shavelson RJ, Webb NM. 1991. Generalizability theory: A primer. In Jaeger RM. Measurement methods for the social sciences series. Sage Publications, INC, pp.102.*
8. *Winterstein BP, Willse JT, Kwapil TR, Silvia PJ. 2010. Assessment of Score dependability of the Wisconsin Schizotypy scales using generalizability analysis. Psychopathol Behav Assess, vol. 32:p.575–585.*
9. *Stora B, Hagtvet KA, Heyerdahl S. 2013. Reliability of observers' subjective impressions of families: A generalizability theory approach. Psychother Res, vol.23:p.448–463.*
10. *Crits-Christoph P, Johnson J, Gallop R, Gibbons MB, Ring-Kurtz S, Hamilton JL, Tu X.2011. A generalizability theory analysis of group process ratings in the treatment of cocaine dependence. Psychother Res, vol.21:p.252–266.*
11. *Heitman RJ, Kovaleski JE, Pugh SF.2009. Application of generalizability theory in estimating the reliability of ankle-complex laxity measurement. J Athl Train, vol.44:p.48–52.*
12. *Cella DF, Tulsky DS, Gray G, Sarafian B, Linn E, Bonomi A.1993. The functional assessment of cancer therapy scale: Development and validation of the general measure. J Clin Oncol, 11(3):570–579.*
13. *Terwee CB, Dekker FW, Wiersinga WM, Prummel MF, Bossuyt PM.2003. On assessing responsiveness of health-related quality of life instruments: guidelines for instrument evaluation. Qual Life Res, vol.12:p.349–363*
14. *Yifang Wu, Hueying Tzou. 2015. A Multivariate Generalizability Theory Approach to Standard Setting. Applied Psychological Measurement, vol.39:p. 507-524.*
15. *Winterstein B P, Willse J T, Kwapil T R, et al. 2010. Assessment of score dependability of the Wisconsin Schizotypy Scales using generalizability analysis. Journal of Psychopathology and Behavioral Assessment, vol. 32:p. 575-585.*