Validation of a Spanish questionnaire implementing the Stanford Educational Framework for Evaluation of Clinical Teachers

Marcela Bitran[a]*, Manuel Torres-Sahli[a], Oslando Padilla[b]

Abstract

Introduction: Although there are instruments in Spanish to evaluate teacher performance during the initial basic science training years or during medical specialization; there are few instruments for the clinical training years, in which the main role of the teacher is to facilitate experiential learning. The MEDUC30 questionnaire is a Spanish instrument developed by the Pontificia Universidad Católica School of Medicine. It was built using the Stanford Faculty Development Program (SFDP) educational framework for evaluation of clinical teachers' effectiveness by students. MEDUC30 has been used since 2004 at Pontificia Universidad Católica de Chile and was previously studied with exploratory methods.

Objective: To provide satisfying evidence of validity and reliability to support MEDUC30's usefulness in Spanish-speaking contexts, using confirmatory analytical methods.

Method: This is an analytical, longitudinal and retrospective study, in which 24,681 MEDUC30 questionnaires evaluating 579 clinical teachers were analysed. They were completed by medical students from 3rd to 7th year of study, from 2004 throughout 2015. The questionnaire's structure was studied by exploratory (EFA) and confirmatory factor analysis (CFA). Measurement invariance was evaluated with multi-group CFA.

Results: Four different models were compared; a bi-factor model was the best alternative to explain the da-

[a]Centro de Educación Médica, Escuela de Medicina, Pontificia Universidad Católica de Chile, Santiago, Chile.
[b]Departamento de Salud Pública, Escuela de Medicina, Pontificia Universidad Católica de Chile, Santiago, Chile.

Recibido: 31-enero-2017. Aceptado: 13-junio-2017.

*Autor de correspondencia: Marcela Bitran, Centro de Educación Médica, Escuela de Medicina, Pontificia Universidad Católica de Chile. Avenida Diagonal Paraguay 362, Santiago, Chile.
Correo electrónico: mbitran@med.puc.cl

La revisión por pares es responsabilidad de la Universidad Nacional Autónoma de México.

https://doi.org/10.1016/j.riem.2017.06.003

ta's structure. It was composed of one general and six domain-specific factors: [i] Patient-Based Teaching, [ii] Communication of Goals, [iii] Evaluation and Feedback, [iv] Promotion of Understanding, Retention, and Self-directed Learning, [v] Control of the Session, and [vi] Learning Climate. The overall reliability of MEDUC30 scores was excellent (Cronbach's α = .98, McDonald's ω = .98) and that of the six specific factors was very good (Cronbach's α =.88-.95, McDonald's ω = .78-.94). Measurement invariance extended over teacher gender, date, semester, year of study, clinical teaching setting, and length of clinical rotation; all of these variables were sources of population heterogeneity.

Conclusions: MEDUC30 is a valid and reliable Spanish instrument to evaluate clinical teachers. It can be used to provide formative feedback to clinical teachers and to provide accurate information to department heads and program directors for resource allocation and promotion purposes.

Keywords: *clinical teacher, evaluation, effectiveness, instrument, and validity.*

Validación de un cuestionario en castellano basado en el Modelo de Stanford para evaluar docentes clínicos

Resumen

Introducción: Aunque existen instrumentos en español para evaluar el desempeño docente durante el ciclo básico o la especialización médica, faltan instrumentos para evaluar la docencia en los años iniciales del entrenamiento clínico, en que el profesor cumple un rol fundamental facilitando el aprendizaje experiencial. MEDUC30 es un cuestionario en español desarrollado por la Escuela de Medicina de la Pontificia Universidad Católica para este efecto. Se construyó con base en el modelo educacional del Programa de Desarrollo Docente de la Universidad de Stanford (SFDP) y ha sido usado desde 2004 en la Pontificia Universidad Católica y validado previamente con métodos exploratorios.

Objetivo: Proveer evidencia de validez y confiabilidad de MEDUC30 que avale su utilidad en contextos hispanohablantes, usando métodos analíticos confirmatorios.

Método: Este es un estudio de carácter analítico, longitudinal y retrospectivo. Se analizaron 24,681 cuestionarios que evaluaban 579 docentes clínicos. Éstos fueron completados por estudiantes de medicina entre tercer y séptimo año, entre 2004 y 2015. Los datos se analizaron mediante análisis factorial exploratorio (AFE) y confirmatorio (AFC), y se evaluó la invariancia de medición con AFC multi-grupo.

Resultados: Se compararon cuatro modelos, de los cuales un modelo bi-factor fue el que mejor dio cuenta de los datos. Este modelo está compuesto de un factor general y seis específicos: [i] Enseñanza centrada en el paciente, [ii] Comunicación de objetivos, [iii] Evaluación y retroalimentación, [iv] Promoción de la comprensión, la retención y el aprendizaje auto-dirigido, [v] Control de la sesión, y [vi] Clima de aprendizaje. La confiabilidad general fue excelente (α Cronbach= .98, ω McDonald = .98) y la de los seis factores, muy buena (α Cronbach =.88-.95, ω McDonald = .78-.94). La invarianza de medición se sostuvo para sexo del docente, fecha, semestre, curso, campo clínico, y duración de la rotación. Todas estas variables mostraron ser fuentes de heterogeneidad poblacional.

Conclusiones: MEDUC30 es un instrumento en español válido y confiable para proveer retroalimentación a los docentes clínicos tanto de su efectividad docente general como de seis dominios educacionales específicos. Además, puede proporcionar información útil para jefes de programas y autoridades para mejorar la calidad de la docencia clínica.

Palabras clave: *Docente clínico, evaluación, calidad, cuestionario, validación.*

INTRODUCTION

In 2000, the School of Medicine of the Pontificia Universidad Católica de Chile created a Faculty Development Center to provide formative instances for its more than 700 clinical teachers, and to foster a culture of continuous enhancement of teaching quality[1,2]. Back then there were few questionnaires in Spanish to evaluate the performance of medical teachers in clinical settings. The Center developed and validated a questionnaire in Spanish named MEDUC30 based on a systematic review of specialized literature and using as a template the seven-domain educational model of Stanford University for evaluating clinical teaching effectiveness[3]. The questionnaire was refined and validated by a Delphi pannel[4].

MEDUC30 is a 30-item questionnaire with a frequency Likert-type scale of 4 points[4]. Its items tribute to one of following domains: [i] Learning Climate, [ii] Evaluation, [iii] Feedback, [iv] Communication of Goals, [v] Control of the Session, [vi] Promotion of Understanding and Retention, [vii] Promotion of Self-directed Learning and[viii] Patient-Based Teaching. The eighth domain was incorporated to ensure that the activity evaluated referred to actual clinical teaching rather than minilectures given in clinical settings.

In the initial validation study[4] MEDUC30 displayed a good reliability and a four-factor structure: Patient-Based Teaching and Learning Climate emerged as separate empirical factors, and the remaining five SFDP's educational domains gathered in two large factors: Evaluating skills (comprising Evaluation plus Feedback) and Teaching skills (comprising Communication of Goals, Control of the Session, Promotion of Comprehension and Retention, and Promotion of Self-directed learning)[4].

MEDUC30 is to our knowledge one of the few validated instrument in Spanish to evaluate clinical teachers' effectiveness during the initial years of clinical experiential learning. It complements the questionnaires developed by the Faculty of Medicine of the UNAM of Mexico to evaluate teachers' performance during the basic medical science teaching[5-7] and medical specialty training[8] and those developed by the Pontificia Universidad Católica de Chile for evaluation of clinical teachers in different medical specialties[9,10]. MEDUC30 has been used at the Pontificia Universidad Católica de Chile medical school since 2004. However, no confirmatory factor analysis (CFA) and measurement invariance studies have been made so far.

The purpose of this study is to provide updated evidence as to the reproducibility, validity and usefulness of MEDUC30 questionnaire to evaluate clinical teaching in Spanish-speaking contexts. To this end, we validated MEDUC30 using a larger and more recent database employing confirmatory analytical methods (CFA) to ascertain the model's goodness-of-fit and multi-group CFA to study measurement invariance.

METHOD

Study and participants

This is an analytical, longitudinal, retrospective study aiming to examine the psychometric properties of the data produced with MEDUC30 in its regular use of evaluation of clinical teachers' performance at Pontificia Universidad Católica de Chile School of medicine. We analysed a total of 24,681 evaluation forms regarding 579 clinical tutors (63% men) collected from 2004 through 2015.

Instrument

MEDUC30 is a 30-item instrument that describes observable teacher behaviours[4]. It uses a four-level scale: 1. 'almost never', 2. 'sometimes', 3. 'often', and 4. 'almost always'[4]. Twenty nine items tribute to eight dimensions as follows: [i] Patient-Based Teaching, items 1-5; [ii] Communication of Goals, items 6-8; [iii] Evaluation, items 9-12; [iv] Promoting Understanding and Retention, items 13-16; [v] Promoting Self-directed Learning, items 17-19; [vi] Control of the Session, items 20-22; [vii] Feedback, items 23-25; and [viii] Learning Climate, items 26-29. The last item corresponds to a global rating.

Application of MEDUC30

The evaluation process was as follows: at the end of each clinical rotation medical students from 3rd to 7th (last) year of study were asked to evaluate their clinical tutors using MEDUC30 paper forms. Students filled the forms anonimoulsy (in the absence of the evaluated teacher) as many times as rotations they had during the year. This was an ongoing process; at the end of each academic year, tutors accumulated between 5 and 30 evaluations. Individual reports were sent to the teachers to provide them specific feedback regarding the eigth domains of effective teaching. Copies of these reports were made available yearly to school authorities to be used as a source of information for academic promotion.

Statistical Analysis

Items as ordinal measures of continuous latent constructs and missing data handling

For analytical purposes, we treated the MEDUC30 items' scores as ordinal rather than continuous[11]. To deal with missing data, we applied multiple imputation[12] using Multivariate Imputation by Chained Equations (MICE) with the proportional odds logistic regression (POLR) model. We used these imputed data to conduct both Exploratory (EFA) and Confirmatory Factor Analyses (CFA).

Factor Analysis

We conducted EFA to study the dimensionality and internal structure of the data, CFA to evaluate the model's goodness-of-fit, and multi-group CFA to study measurement invariance.

The data was randomly divided into three samples: sample 1 (n = 4,122) for EFA, sample 2 (n = 4,109) for CFA global fit assessment, and sample 3 (n = 16,450) for CFA measurement invariance evaluation.

For the EFA we used the unweighted least squares (ULS) estimator, while for CFA, we added robust standard errors, and mean and variance adjusted test statistic with second order approach[13] (ULSMVS in lavaan R Package).

Four fit indices were calculated to evaluate and compare descriptive goodness-of-fit. Two comparative fit indices: Comparative Fit Index (CFI) and Tucker-Lewis Index (TLI); one parsimony correction index: Root-Mean-Square Error of Approximation (RMSEA); and one absolute fit index: Weighted Root-Mean-square Residual (WRMR).

The following cutoff values were derived from simulation studies[14-17]: good fit when: CFI $\geq .96$, TLI $\geq .95$, RMSEA $\leq .05$; acceptable fit when: CFI and TLI $\geq .90$, RMSEA $< .08$; mediocre fit: if $.08 \leq$ RMSEA $\leq .10$, with CFI and TLI $\geq .90$. Meeting at least two of the three criteria just described in one level of satisfaction, and the remaining in an adjacent level (upper or lower), the model fit was assumed as conforming to the former[16,18]. Finally, if CFI or TLI $< .90$, or RMSEA $> .10$ the model were rejected. WRMR (smaller is better) was used to corroborate model comparison[19].

For reliability measures, we calculated Cronbach's α, and McDonald's ω and ω_t[20] as indices of internal consistency of the respective constructs. Additionally, for bifactor model constructs we reported McDonald's ω_h[21] with respective ω_s coefficients as indices of factor saturation.

We studied measurement invariance over tutor gender, date, semester, year of study, clinical teaching setting, and length of clinical rotation using χ^2-based likelihood-ratio test (LRT) with Satorra[22] adjusted test statistic. For every grouping variable, we used a random subsample of the biggest group(s) to ensure equal *n* with the smaller ones.

Software

We conducted all statistical analyses using R software 3.3.0 [23] with specific packages. Multiple Imputation by Chained Equations was performed using the *mice* package 2.25 [24]. Exploratory factor analysis, including tests for sampling adequacy and parallel analysis, was conducted with the *psych* package 1.6.4 [25]. Confirmatory Factor Analysis was conducted with the lavaan package 0.5.20 [26]. Proportional odds regressions were performed with the MASS package 7.3.45 [27].

Ethical Considerations

This paper reports the results of the clinical teachers' evaluations conducted from 2004 to 2015 at Pontificia Universidad Católica de Chile School of Medicine. This is a mandatory process overseen by the Center for Medical Education according to ethical considerations aimed to assure the confidential handling of the information. The evaluation forms are filled anonymously by students. Each clinical tutor receives a yearly report of his/her results and these results are also known by department head and the school authorities to be used for purposes of career promotion.

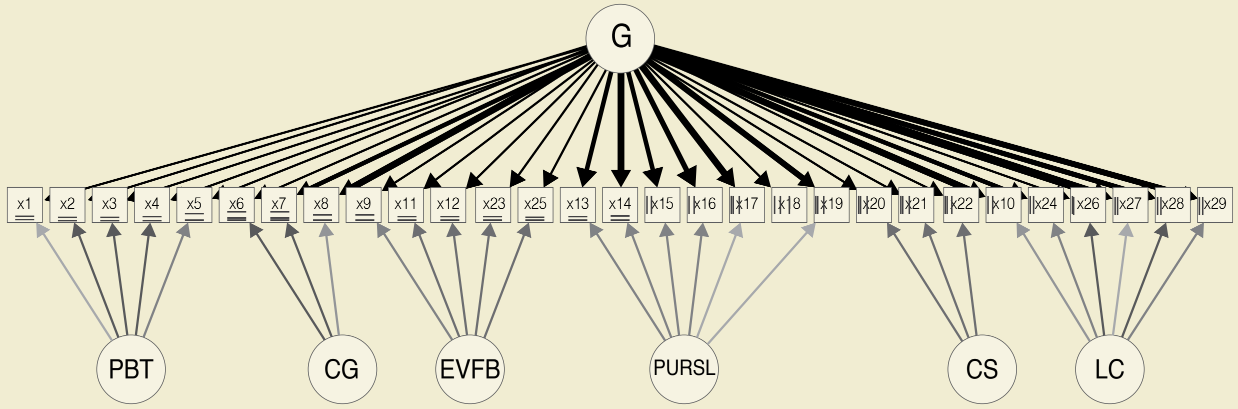

Figure 1. Diagram of the bi-factor model of MEDUC30 scores

RESULTS

The proportion of missing responses for the whole questionnaire was low (1.79%) except for items 10 and 24 (8.72% and 5.13% missing values, respectively). To deal with this situation, we used multiple imputation.

On the other hand, the highest response category of the questionnaire ('almost always') concentrated 74.3% of the answers, indicative of a 'ceiling effect'. To deal with this situation, we selected polychoric correlation and a robust estimator[19,28,29].

The three samples were comparable with regards to date (ANOVA p = .60), tutor gender, clinical teaching setting, semester (binomial regression p = .29, .72, and .43 respectively), year of study, and global score (POLR p = .33, .69 respectively).

Exploratory Factor Analysis (Sample 1)

Assumptions' Evaluation

The Kaiser-Meyer-Olkin measure of sampling adequacy (.97) indicated a *marvellous* (≥ .90) factorability according to Kaiser and Rice[30] criteria. The Bartlett's sphericity test (^{2} = 129511.72, df = 406, p < .001) also indicated reasonability of factor analysis.

Number of Retained Factors

We decided to retain seven factors according to two criteria: the maximum number of factors given by Horn's Parallel Analysis (eight factors) and the last big drop in the eigenvalues in the sedimentation plot (between factors 7 and 8).

In addition, we tested the hypothesis that a bifactor model could better explain the data given the large difference in eigenvalues between factors 1 and 2 (17.36 vs. 0.98), as suggested by Reise[31]. This implied the retention of one general factor, with six specific factors.

Latent Structure

Exploratory analysis for a bifactor structure model with 6 specific dimensions resulted in the following domain-specific factors, alongside the general factor (**figure 1**): Patient-Based Teaching (PBT; items 1 to 5), Communication of Goals (CG; items 6 to 8), Evaluation and Feedback (EVFB; items 9, 11, 12, 23, and 25), Promotion of Understanding, Retention and Self-directed Learning (PURSL; items 13 to 19), Control of the Session (CS; items 20 to 22), and Learning Climate (LC; items 10, 24, and 26 to 29). Only two items had considerable cross-loadings (items 10 and 24), both loaded more on Learning Climate than on their original theoretical factor. Commonalities of items ranged from .50 to .90. Factor loadings ranged from .59 to .83 for the general factor, and from .23 to .67 for domain-specific factors.

Confirmatory Factor Analysis (Sample 2)

Goodness-of-fit and model comparison

We compared four models in CFA: [i] The four correlated traits model described by Bitran et al.[4], [ii] the bifactor model suggested in the results of EFA, [iii]

Table 1. CFA Global goodness-of-fit indices (Sample 2)

Model	$SB\chi^2$ (*df*)	$SB\chi^2$ / *df*	CFI	TLI	RMSEA [90% CI]	WRMR
Sample 2						
Single factor	4029.6 (154)	26.1	.771	.979	.078 [.076, .080]	3.49
Four Correlated Traits	2531.5 (169)	15.0	.86	.988	.058 [.057, .060]	2.57
Six Correlated Traits	2029.2 (176)	11.6	.89	.991	.051 [.049, .053]	2.17
Bifactor	1234.2 (3)	8.1	.936	.994	.041 [.040, .043]	1.82

Note. $SB\chi^2$ = Satorra-Bentler scaled chi-square, *df* = degrees of freedom, CFI = Comparative fit index. TLI = Tucker-Lewis index. RMSEA = Root-mean-square error of approximation. WRMR = Weighted root-mean-square residual. All *p*-values < .001.

a model with six correlated traits (corresponding to the six domain-specific factors of the bifactor model), and [iv] a single-factor model. According to evaluated global fit indices **(table 1)**, the bifactor model was the only one with acceptable (CFI) to good (TLI and RMSEA) global fit indices. This supports the bifactor model with six domain-specific factors as the best latent structure for this MEDUC30 Data.

Factor structure and reliability

All factors in the bifactor model (the general and the six domain-specific factors) displayed good reliability coefficients: .88 to .98 for Cronbach's α, and .79 to .97 for McDonald's ω (Table 2). Hierarchical reliability ($\omega_{h/s}$) was stronger for the general factor compared to the domain-specific factors, particularly the PURSL factor (ω_s= .08) (Table 2).

Factor loadings for the bifactor model **(table 2)** were all high on the general factor, ranging from .63 (item 22) to .87 (item 27). All specific factors except PURSL had at least two salient loadings (≥ .40). With exception of items 1, 8, 17, 18, 19 and 27 (loading < .20), domain-specific loadings were in general large enough (≥ .20) reflecting a multidimensional structure.

Measurement invariance (Sample 3)

Multigroup CFA **(table 3)** indicated that configural (form), weak (loadings) and strong (intercepts) measurement invariance could be sustained (p > .05) across tutor gender (man/woman), clinical teaching setting (inpatient/outpatient), year of study (3rd to 7th), length of clinical rotation (1 to 7-or-more weeks), date (2004 to 2015), and semester (fall/spring). Also, all of these variables were sources of population heterogeneity (mean) with p <.001 except for the length of clinical rotation with p = .003 which was still significant at most traditional confidence values.

Discussion

We evaluated the validity of MEDUC30 to assess clinical teachers' effectiveness. According to EFA, our data was reasonably well explained by a bifactor structure with six domain-specific factors. Five of them closely related to the seven theoretical domains of SFDP framework[3], and the sixth factor corresponding to an added dimension named Patient-Based Teaching. CFA proved that this model had a good fit for the data and was better than a single factor model or a first-order multidimensional model with four or six factors to account for MEDUC30 scores.

Besides supporting the multidimensionality of the teaching effectiveness construct, present results indicate that MEDUC30 behaves as a hierarchical construct, with a general factor that can be construed as 'being a good teacher', and six domain-specific factors. During the last decade, hundreds of clinical teachers at the PUC have completed a diploma in medical education[32]. Thus, it is conceivable that the 'being a good teacher' general factor found in this study is related to this professionalisation of teaching which entails the acquisition of general good teaching practices, in addition to domain-specific skills.

The 'good teacher' general factor found in our study is reminiscent of the 'teaching performance' latent construct proposed by Flores et al.[7] to explain his results with OPINEST, an instrument used to evaluate medical teacher competences.

Table 2. CFA Standardized factor loadings and strength indices for bifactor model (Sample 2)

Item	Theoretical factor	Bifactor Model (general & six domain-specific factors)						
		g	PBT	CG	EVFB	PURSL	CS	LC
Item 1	PBT	.71	.09					
Item 2	PBT	.71	.60					
Item 3	PBT	.73	.59					
Item 4	PBT	.70	.55					
Item 5	PBT	.73	.28					
Item 6	CG	.76		.53				
Item 7	CG	.80		.51				
Item 8	CG	.83		.16				
Item 9	EV	.71			.35			
Item 10	EV	.83						.21
Item 11	EV	.70			.42			
Item 12	EV	.80			.45			
Item 13	PUR	.80				.27		
Item 14	PUR	.82				.35		
Item 15	PUR	.81				.36		
Item 16	PUR	.82				.29		
Item 17	PSL	.82				.13		
Item 18	PSL	.72				.03		
Item 19	PSL	.82				.15		
Item 20	CS	.76					.47	
Item 21	CS	.73					.49	
Item 22	CS	.63					.42	
Item 23	FB	.73			.41			
Item 24	FB	.81						.29
Item 25	FB	.74			.40			
Item 26	LC	.78						.49
Item 27	LC	.89						.15
Item 28	LC	.80						.48
Item 29	LC	.85		.89				.40
Cronbach's α		.98	.91	.91	.92	.94	.88	.96
McDonald's ω_t		.97	.86	.87	.87	.88	.79	.93
McDonald's $\omega_{(h/s)}$		.92	.31	.21	.23	.08	.21	.25

Note. g = General factor, PBT = Patient-based teaching, CG = Communication of goals, EVFB = Evaluation and feedback, PURSL = Promotion of understanding, retention and self-directed learning, CS = Control of session, LC = Learning Climate.

The potential contribution of MEDUC30 to medical education in Spanish-speaking contexts is related to its focus on the evaluation of facilitatory rol of medical teachers in the clinical setting. MEDUC30 covers this sensitive period of transition from passive, teacher-centered, information-driven teaching to active, student-centered, patient-driven learning[33]. A recent study found that the attributes of an effective teacher differ between the classroom and the clinical setting[34] thus giving support to the importance of context specificity in teaching effectiveness ratings.

Table 3. Likelihood-ratio test (χ^2 difference test) for Multi-group Measurement Invariance (Sample 3)

Invariance level	**Bifactor model**			
	χ^2 (*df*)	**$\Delta\chi^2$ [CFI]**	**Δ*df* [RMSEA]**	***p* (>χ^2)**
***Gender*[a]**				
Configural	3485 (689)	[.98]	[.034]	
Loadings	3870 (747)	3.7	3.15	.32
Intercepts	4058 (798)	3.7	8.28	.90
Means	5156 (805)	78.8	5.13	<.001
***Semester*[b]**				
Configural	3547 (689)	[.98]	[.035]	
Loadings	4033 (747)	4.5	3.01	.22
Intercepts	3996 (798)	-0.8	9.00	>.999
Means	4516 (805)	38.0	5.37	<.001
***Clinical teaching setting*[c]**				
Configural	3997 (689)	[.98]	[.034]	
Loadings	4946 (747)	7.1	2.648	.051
Intercepts	5195 (798)	4.5	8.23	.827
Means	6954 (805)	124.1	5.29	<.001
***Year of study*[d]**				
Configural	2928 (1712)	[.98]	[.032]	
Loadings	5227 (1944)	11.0	5.63	.074
Intercepts	5078 (2148)	-0.6	7.27	>.999
Means	11544 (2176)	107.4	5.00	<.001
***Length of clinical rotation*[e]**				
Configural	3818 (2394)	[.98]	[.032]	
Loadings	5860 (2742)	6.08	5.40	.345
Intercepts	6678 (3048)	2.35	7.45	.954
Means	8226 (3090)	17.55	4.79	.003
***Longitudinal*[f]**				
Configural	5324 (4099)	[.98]	[.032]	
Loadings	9078 (4737)	6.78	5.94	.34
Intercepts	9756 (5298)	0.99	7.01	>.999
Means	15012 (5375)	30.28	4.58	<.001

Note. SBχ^2 = Satorra-Bentler scaled chi-square, *df* = degrees of freedom. CFI = Comparative fit index, configural invariance only. RMSEA = Root-mean-square error of approximation, configural invariance only. [a] n = 4000 per gender of the tutor (man/woman). [b] n = 4000 per semester (fall/spring). [c] n = 4000 per clinical teaching setting (inpatient/outpatient). [d] n = 933 per year of study (3rd, 4th, 5th, 6th, 7th). [e] n = 780 per group (one, two, three, four, five, six, or seven-and-more weeks). [f] n = 503 per year (2004 to 2015).

Similarities and differences with other implementations of the SFDP framework

Our results are partially consistent with initial validations of the SFDP framework construct using EFA on data obtained with the questionnaire SFDP26[3,35]. In these studies the authors deemed the data to be reasonably explained by the theoretical seven-dimension structure.

Compared to the four-factor structure proposed for MEDUC30 in the initial exploratory studies[4], the bifactor structure with six domain-specific factors presented here corresponds more closely to the theoretical SFDP framework. Three of these factors corresponded exactly to the dimensions: Learning Climate, Control of the Session and Communication of Goals. The other two factors gathered the items of Evaluation and Feedback, on the one hand, and Promoting Comprehension and Retention and Promoting Self-directed Learning, on the other.

In a psychometric evaluation of the SFDP26, performed over a relatively small sample (N = 119), Mintz et al.[36] proposed a new five-factor structure for a reduced 15-item instrument. Comparisons of our results with this report are difficult to draw since the authors did not evaluate hierarchical models. On the other hand, they eliminated entire dimensions rather than redefining the structure based on substantive and statistical criteria with the original set of items.

In a recent study done with Middle Eastern undergraduate medical students[37], a modified version of the "System for Evaluation of Teaching Qualities (SETQ), an instrument also based on the SPDF educational famework, displayed a six-factor structure consistent with the main SFDP domains and with MEDUC30.

Strengths and limitations of this study

MEDUC30 is a validated theory-based instrument in Spanish to assess clinical teachers' effectiveness by students during the training clinical years. It adds to the repertoire of instruments developed in Spanish to evaluate medical teacher performance in basic science years[5,6], and those aimed at medical specialty training[8-10].

This study has strengths related to the sample size and analytical methods used. Compared to other validation studies (for a revision see Fluit et al.[2]), the number of evaluations and teachers was several folds larger, and we employed multiple and stringent criteria for the factor analyses. These features endorse the robustness and reliability of results.

Unlike most validations of similar instruments, this study includes measurement invariance information. MEDUC30 can be used for comparisons across several data variables (i.e. date, tutor gender, year of study, and the length of rotation).

One limitation of this study is that it involves a single medical school; thus it would be necessary to confirm the questionnaire generalizability for other medical schools or countries with different clinical teaching realities.

Regarding further improvements of the questionnaire, it seems advisable to increase the width of the scale to allow for a larger response range. Scores have improved systematically during the 12-year period of assessment and, as a consequence, the power of discrimination of the 4-point scale has diminished.

It should always be borne in mind that while the assessment of clinical teachers by students could reveal valid and relevant information, this should be "triangulated" with information derived from other sources, including peers and self-assessment[2].

CONCLUSIONS

In this report we give evidence that MEDUC30 is a reliable and valid instrument suited to provide clinical teachers with feedback on their strengths and weaknesses about multiple dimensions of clinical teaching. It has evidence of content validity, internal structure validity and use validity. This instrument should be of interest to medical schools of Spanish-speaking countries for it adds to the repertoire of validated instruments in Spanish to evaluate medical teachers in the clinical teaching setting.

MEDUC30 internal structure validity was supported in this study by the multidimensionality of its scores and the consistency of this internal structure with the educational framework used in its development. The content validity evidence derives from the questionnaire construction, which was based on a previously developed instrument, and on the input of experts and students. Also, MEDUC30 items cover 5 out of 7 of the roles agreed as characteristic of good teaching[38-40]. Finally, its use validity is confirmed by the widespread use and acceptance of this instrument for the assessment of clinical teachers at PUC medical school for more than ten years.

In conclusion, MEDUC30 meets satisfactorily three of the five possible sources of validity evidence as defined by the American Psychological and Education Research Associations published standards[41,42]: internal structure, content and use. Future investigations will be needed to provide evidence for the remaining two validity sources: relation to other variables and consequences.

CONTRIBUCIÓN INDIVIDUAL

- **MB:** Concepción y diseño del estudio, interpretación de los datos, discusión, redacción del manuscrito y aprobación final del mismo para envío a publicación.
- **MT-S:** Análisis e interpretación de la data, elaboración de un manuscrito inicial, discusión y análisis de resultados, revisión crítica del manuscrito final.
- **OP:** Análisis e interpretación de la data, discusión y análisis de resultados, revisión crítica del manuscrito final.

AGRADECIMIENTOS

Agradecemos a todos los estudiantes y docentes que participaron del proceso de evaluación durante el período 2004-2015.

PRESENTACIONES PREVIAS

Ninguna.

FINANCIAMIENTO

Reconocemos el apoyo del Centro de Educación Médica de la Escuela de Medicina de la Pontificia Universidad Católica de Chile.

CONFLICTOS DE INTERÉS

Ninguno.

REFERENCES

1. Snell L, Tallett S, Haist S, Hays R, Norcini J, Prince K, et al. A review of the evaluation of clinical teaching: new perspectives and challenges. Med Educ [Internet]. 2000 Oct;34(10):862–70. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11012937
2. Fluit C, Bolhuis S, Grol R, Laan R, Wensing M. Assessing the quality of clinical teachers: a systematic review of content and quality of questionnaires for assessing clinical teachers. J Gen Intern Med [Internet]. 2010 Dec;25(12):1337–45. Available from: http://dx.doi.org/10.1007/s11606-010-1458-y
3. Litzelman DK, Stratos GA, Marriott DJ, Skeff KM. Factorial validation of a widely disseminated educational framework for evaluating clinical teachers. Acad Med [Internet]. 1998 Jun;73(6):688-95. Available from: http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage&an=00001888-199806000-00016
4. Bitran M, Mena B, Riquelme A, Padilla O, Sánchez I, Moreno R. An instrument in Spanish to evaluate the performance of clinical teachers by students. Revista Médica de Chile [Internet]. 2010;138(6):685–93. Available from: http://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0034-988720100006000045.
5. Valle R, Alaminos I, Contreras E, Salas L, Tomasini P, Varela M. Student Questionnaire to evaluate basic medical science teaching (METEBQ-B). Rev Med IMSS 2004;42(5):405-411. Available from: http://www.facmed.unam.mx/sem/pdf/articulosrocio/StudentQuestionnaire.pdf
6. Mazón J, Martínez J, Martínez A. La evaluación de la función docente mediante la opinión del estudiante. Un nuevo instrumento para nuevas dimensiones: COED. RESU. 2009;38:113-140.
7. Flores F, Gatica F, Sánchez-Mandiola M, Martínez A. Evolución de la evaluación del desempeño docente en la Facultad de Medicina; evidencia de validez y confiabilidad. 2017 Inv Ed Med 6(22)96-103.
8. Martínez A, Lifshitz A, Ponce R, Aguilar V. Evaluación del desempeño docente en cursos de especialización médica; validación de un cuestionario. 2008 Rev Med IMSS;46:375-382.
9. Pizarro M, Solis N , Rojas V, Diaz L, Padilla O, Letelier L, Aizman A, Sarfatis A, Olivos T, Soza A, Delfino A, Latorre G, Ivanovic D , Hoyl T, Bitran M, Arab J, Riquelme A. Development of MEDUC-PG14 survey to assess postgraduate teaching in medical specialties. 2015 Revista Medica De Chile;143(8):1005-1014.
10. Huete A, Julio R, Rojas V, Herrera C, Padilla O, Solís N, Pizarro M, Etcheberry L, Sarfatis A, Pérez G, Delfino A, Muñoz E, Rivera H, Bitrán M, Riquelme A. Development and validation of the MEDUC-RX32 questionnaire, to evaluate teachers of postgraduate radiology programs. 2014 Revista Chilena De Radiología;20(2):75-80.
11. Rubin DB. Multiple Imputation after 18+ Years. J Am Stat Assoc [Internet]. 1996;91(434):473–89. Available from: http://www.tandfonline.com/doi/abs/10.1080/01621459.1996.10476908
12. Rhemtulla M, Brosseau-Liard PÉ, Savalei V. When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. Psychol Methods [Internet]. 2012 Sep;17(3):354–73. Available from: http://dx.doi.org/10.1037/a0029315
13. Rubin DB. Multiple Imputation after 18+ Years. J Am Stat Assoc [Internet]. 1996;91(434):473–89. Available from: http://www.tandfonline.com/doi/abs/10.1080/01621459.1996.10476908
14. Bentler PM. Comparative fit indexes in structural models. Psychol Bull [Internet]. 1990 Mar;107(2):238–46. Available from: http://www.ncbi.nlm.nih.gov/pubmed/2320703
15. Browne MW, Cudeck R. Alternative Ways of Assessing Model Fit. Sociol Methods Res [Internet]. 1992 Nov 1;21(2):230–58. Available from: http://smr.sagepub.com/content/21/2/230.abstract
16. Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Struct Equ Modeling [Internet]. 1999;6(1):1–55. Available from: http://dx.doi.org/10.1080/10705519909540118
17. Yu C-Y. Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes [Inter-

net]. University of California Los Angeles; 2002. Available from: http://ww.statmodel2.com/download/Yudissertation.pdf

18. Brown TA. Confirmatory factor analysis for applied research. Second. New York: The Guilford Press; 2015.
19. Muthén LK, Muthén BO. Mplus User's Guide. Seventh. Los Angeles, CA; 2012.
20. McDonald RP. Test Theory: A Unified Treatment [Internet]. Psychology Press; 2013. 498 p. Available from: http://books.google.cl/books/about/Test_Theory.html?hl=&id=2-V5tOsa_DoC
21. Zinbarg RE, Revelle W, Yovel I, Li W. Cronbach's α, Revelle's β, and McDonald's ω H: their relations with each other and two alternative conceptualizations of reliability. Psychometrika [Internet]. 2005 Apr 2 [cited 2016 Apr 27];70(1):123–33. Available from: http://link.springer.com/article/10.1007/s11336-003-0974-7
22. Satorra A. Scaled and Adjusted Restricted Tests in Multi-Sample Analysis of Moment Structures. In: Heijmans RDH, Pollock DSG, Satorra A, editors. Innovations in Multivariate Statistical Analysis [Internet]. Boston, MA: Springer US; 2000 [cited 2016 May 22]. p. 233–47. (Advanced Studies in Theoretical and Applied Econometrics; vol. 36). Available from: http://link.springer.com/chapter/10.1007/978-1-4615-4603-0_17
23. R Core Team. R: A language and environment for statistical computing [Internet]. Vienna, Austria.: R Foundation for Statistical Computing; 2016. Available from: https://www.r-project.org/
24. van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. J Stat Softw [Internet]. 2011;45(3):1–67. Available from: http://www.jstatsoft.org/v45/i03/
25. Revelle W. psych: Procedures for Psychological, Psychometric, and Personality Research [Internet]. Evanston, Illinois: Northwestern University; 2016. Available from: http://CRAN.R-project.org/package=psych
26. Rosseel Y. lavaan: An R Package for Structural Equation Modeling. J Stat Softw [Internet]. 2012;48(1):1–36. Available from: https://www.jstatsoft.org/index.php/jss/article/view/v048i02
27. Venables WN, Ripley BD. Modern Applied Statistics with S [Internet]. Fourth. New York: Springer; 2002. 495 p. (Statistics and Computing). Available from: http://books.google.cl/books/about/Modern_Applied_Statistics_with_S.html?hl=&id=GeX9CYd_JTkC
28. Jöreskog KG. Censored variables and censored regression [Internet]. 2002 [cited 2016 May 10]. Available from: http://www.ssicentral.com/lisrel/techdocs/censor.pdf
29. Kline RB. Principles and practice of structural equation modeling. Fourth. New York: The Guilford Press; 2015.
30. Kaiser HF, Rice J. Little Jiffy, Mark IV. Educ Psychol Meas [Internet]. 1974 [cited 2016 May 11]; Available from: http://dx.doi.org/10.1177/001316447403400115
31. Reise SP. Invited Paper: The Rediscovery of Bifactor Measurement Models. Multivariate Behav Res [Internet]. 2012 Sep 1;47(5):667–96. Available from: http://dx.doi.org/10.1080/00273171.2012.715555
32. Triviño X, Ximena T, Marisol S, Philippa M, Luz M. Impacto de un programa de formación en docencia en una escuela de medicina [Impact of a diploma on medical education in a medical school in Chile]. Revista médica de Chile [Internet]. 2011;139(11):1508–15. Available from: http://dx.doi.org/10.4067/s0034-98872011001100019
33. Bitran M, Zúñiga D, Pedrals N, Padilla O, Mena B. Medical students' change in learning styles during the course of the undergraduate program: from 'thinking and watching' to 'thinking and doing'. 2012 Canadian Medical Education Journal 2012, 3(2)e86:e97
34. Haws J, Rannelli L, Schaefer JP, Zarnke K, Coderre S, Ravani P, McLaughlin K. The attributes of an effective teacher differ between the classroom and the clinic setting. 2016. Adv in Health Sci Educ (2016) 21:833–840 DOI 10.1007/s10459-016-9669-6
35. Litzelman DK, Westmoreland GR, Skeff KM, Stratos GA. Factorial validation of an educational framework using residents' evaluations of clinician-educators. Acad Med [Internet]. 1999 Oct;74(10):S25–7. Available from: http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage&an=00001888-199910000-00030
36. Mintz M, Southern DA, Ghali WA, Ma IWY. Validation of the 25-Item Stanford Faculty Development Program Tool on Clinical Teaching Effectiveness. Teach Learn Med [Internet]. 2015;27(2):174-81. Available from: http://dx.doi.org/10.1080/10401334.2015.1011645
37. Al Ansari, A., Strachan, K., Hashim, S., & Otoom, S. Analysis of psychometric properties of the modified SETQ tool in undergraduate medical education. 2017 BMC Medical Education. 17, 56. http://doi.org/10.1186/s12909-017-0893-4
38. AMEE Guide No 20: The good teacher is more than a lecturer: the twelve roles of the teacher R.M. HARDEN & JOY CROSBY Centre for Medical Education, University of Dundee, UK. 2000 Medical Teacher 22(4):334-347
39. Boor K, Teunissen PW, Scherpbier AJJA, van der Vleuten CPM, van de Lande J, Scheele F. Residents' perceptions of the ideal clinical teacher—A qualitative study. Eur J Obstet Gynecol Reprod Biol [Internet]. 2008 Oct;140(2):152–7. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0301211508001395
40. Paukert JL, Richards BF. How medical students and residents describe the roles and characteristics of their influential clinical teachers. Acad Med [Internet]. 2000 Aug;75(8):843-5. Available from: http://www.ncbi.nlm.nih.gov/pubmed/10965865
41. Beckman TJ, Cook DA, Mandrekar JN. What is the validity evidence for assessments of clinical teaching? J Gen Intern Med [Internet]. 2005 Dec;20(12):1159–64. Available from: http://dx.doi.org/10.1111/j.1525-1497.2005.0258.x
42. Downing SM. Reliability: on the reproducibility of assessment data. Med Educ [Internet]. 2004 Sep;38(9):1006–12. Available from: http://dx.doi.org/10.1111/j.1365-2929.2004.01932.x