Anuario de Psicología
The UB Journal of Psychology | 53/2
Universitat de Barcelona

AUTORES

Alejandro Veas
Departamento de Psicología Evolutiva y de la Educación. Universidad de Murcia, España
orcid.org/0000-0002-5560-2215
alejandro.veas@um.es

Leandro Navas
Departamento de Psicología Evolutiva y Didáctica. Universidad de Alicante, España
doi.org/0000-0003-3005-9073
leandro.navas@um.es

Anuario de Psicología
N.º 53/2 | 2023 | págs. 23-32

Enviado: 10 de noviembre de 2022
Aceptado: 2 de marzo de 2023

DOI: 10.1344/ANPSIC2023.53/2.3

ISSN: 0066-5126 |

Development and psychometric properties of the Student Evaluation of Dissertation Tutoring Scale (SEDITUS) in higher education

Alejandro Veas y Leandro Navas

Abstract

Academic dissertation tutoring has become an important academic teaching competence in higher education. However, no assessment instrument has been developed to measure the needed transversal competences in quality tutoring. In two studies, we aim to develop and explore the psychometric properties of the Student Evaluation of Dissertation Tutoring Scale (SEDITUS). In Study 1 (N = 82, 72% women), the initial eight items proposed a unidimensional construct and good fit properties to the Polytomous Rasch model. Differential item functioning (DIF) revealed no significant differences across gender for all the items, and the four-category structure did function well. In Study 2 (N = 1046, 69% women), an expert committee decided to remove one item due to the lack of a generalization process for the item across faculty degrees. A multilevel Rasch model was used to consider the nested nature of the data (students nested in eight faculties). Results replicated Study 1, with the scale showing appropriate psychometric properties at the item and global level. Overall, the two studies suggest the SEDITUS can be recommended as a rapid assessment of the tutoring process in dissertations developed both in undergraduate and master's degrees.

Keywords

Higher education, academic dissertations, dissertation tutoring, Rasch model, polytomic scale.

Desenvolupament i propietats psicomètriques de l'escala d'avaluació de la tutorització de treballs acadèmics (SEDITUS) en l'Educació Superior

Resum

Els processos de tutorització de treballs acadèmics impliquen una competència important en el professorat universitari. Tot i això, no hi ha escales capaces de mesurar les destreses associades a la qualitat de la tutorització. A partir de dos estudis empírics, es pretén desenvolupar i avaluar les propietats psicomètriques de l'Escala d'Avaluació estudiantil dels processos de tutorització de treballs (SEDITUS). En l'estudio 1 (N = 82, 72% dones), es van proposar 8 ítems inicials ajustats a un model unidimensional a partir del model de Rasch politòmic. Les anàlisis de funcionament diferencial dels ítems (FID) no van mostrar diferències significatives de gènere en cap ítem, i l'estructura de 4 categories va mostrar un rendiment adequat. En l'estudio 2 (N = 1046, 69% dones), un comitè d'experts va decidir eliminar un ítem perquè era incapaç de generalitzar el procés descrit en les titulacions associades a les facultats. Es va emprar un model de Rasch multinivell que tenia en compte l'estructura niada de les dades (estudiants niats de 8 facultats). Els resultats van replicar els del primer estudi i mostren propietats psicomètriques apropiades a nivell d'ítem i de constructe. En conjunt, tots dos estudis suggereixen que l'escala SEDITUS és recomanable per mesurar ràpidament els processos de tutorització de treballs acadèmics.

Paraules clau

Educació superior; treballs acadèmics; tutorització de treballs; model de Rasch; escala politòmica.

Desarrollo y propiedades psicométricas de la escala de evaluación de la tutorización de trabajos académicos (SEDITUS) en Educación Superior

Resumen

Los procesos de tutorización de trabajos académicos suponen una competencia importante en el profesorado universitario. Sin embargo, no se disponen de escalas capaces de medir las destrezas asociadas a la calidad de la tutorización. Partiendo de 2 estudios empíricos, se pretende desarrollar y evaluar las propiedades psicométricas de la Escala de Evaluación estudiantil de los procesos de tutorización de trabajos (SEDITUS). En el estudio 1 (N = 82, 72% mujeres), se propusieron 8 ítems iniciales ajustaron a un modelo unidimensional a partir del modelo de Rasch politómico. Los análisis de funcionamiento diferencial de los ítems (FID) no mostraron diferencias significativas de género en ningún ítem, y la estructura de 4 categorías mostró un adecuado rendimiento. En el estudio 2 (N = 1046, 69% mujeres), un comité de expertos decidió eliminar un ítem debido a su incapacidad de generalización del proceso descrito en las titulaciones asociadas a las facultades. Se empleó un modelo de Rasch multinivel que tuvo en cuenta la estructura anidada de los datos (estudiantes anidados en 8 facultades). Los resultados replicaron los del primer estudio, mostrando propiedades psicométricas apropiadas a nivel de ítem y de constructo. En conjunto, ambos estudios sugieren que la escala SEDITUS es recomendable para una medición rápida de los procesos de tutorización de trabajos académicos.

Palabras clave

Educación superior; trabajos académicos; tutorización de trabajos; modelo de Rasch; escala politómica.

INTRODUCTION

Academic dissertations are considered as one of the most important learning tools, which in fact constitutes the final assessment procedure for undergraduate and master's degree projects in most of the official teaching qualifications in higher education. Similarly, assessment processes have moved towards a dynamic system wherein both professors and students are considered the main agents. Considering an ecological learning perspective (Saroyan & Amundsen, 2001), the present research aims to measure students' satisfaction level with respect to the monitoring process developed by their tutors at their final master's/degree dissertation. To this purpose, this type of scale may provide basic contents by which tutors can reflect over the implemented competences in these subjects.

During recent years, there has been an increase of teaching-learning assessment programmes with a focus on improving both students' implication and teachers' efficacy levels (Webb & Jones, 2009; Wyatt-Smith et al., 2010). In the context of higher education, these programmes consider the professional competences framework, defined as the set of abilities, attitudes and responsibilities that describe learning results in an educational process (Stierer & Antoniou, 2004). In this field, evaluation of teaching has become an extensive practice in higher education (Huybers, 2014; Richardson, 2005). This practice is understood as an assessment of elements associated with professors' achievement in their respective subjects, and is used with promotion purposes, such as professional enrolment and upgrades (Linse, 2017).

During recent decades, multiple explanatory models of teaching process have been developed. Initially, teaching competences were only focused under the behaviourist perspective (Boice, 1991), and secondly to complex cognitive and affective activities (Leinhardt & Greeno, 1986).

Considering these structures, the initial models come from the expertise or ability theories, where the main indicators are based on the organization of knowledge and its structure, the efficacy of procedures and meta-cognitive abilities, among others (Glaser et al., 1984). However, and especially in higher education, beliefs or views about teaching were also considered as relevant for effective teaching (Larsson, 1986). Considering beliefs as implicit assumptions about students, classroom learning and subject knowledge (Pajares, 1992; Pratt, 1997), this model evolved into a more integrated teaching-learning framework, using both constructs to justify an effective teaching assessment.

From a more dynamic perspective, transformative learning theory (Mezirow, 1991) declared an important difference between individuals' prior beliefs about teaching and real performance during the teaching process. In these situations, the identification of contextual barriers is essential, together with those professors with a lack of effective teaching strategies in educational settings (Schön, 1983). The change to effective teaching is based on an engagement process for giving support and positive strategies. Within this line, remarkable is the concept of assessment literacy (Stiggins, 1991), in which professors' assessment knowledge and practice abilities are considered for the interpretation and design of instruction and feedback assignment.

Considering current contextual perspectives, the authors of this study rely on the ecological model proposed by Saroyan & Amundsen (2001), which considers three main elements associated with teaching assessment: conceptions or beliefs, knowledge, and actions. In addition, the influence of context is considered as the key for a dynamic conception of these constructs, as external factors – university/faculty/department culture or professors' specific tasks – may have an impact on teaching tasks (Saroyan & Amundsen, 2001).

The international literature considers dissertations as an important learning process in many universities. According to Evans et al. (2020), in the last 20 years there has been a significant increase of this pedagogic works among disciplines and countries. However, there are some concerns about the way pedagogical research is transferred into educational practice within each branch of knowledge (Stierer & Antoniou, 2004). Following these lines, the concept of world standard (AERA, 2018) refers to those which stand out by their originality, significance, rigorousness, and relevance, with no mention of geographic nature or other particular subjects' background factors. The way individuals and community fields interpret joint and separate terms and definitions provides complex holistic judgements that may disserve the application in the assessment field.

Scientific literature has shown a wide variety of scales for the evaluation of teaching, and most of them consider the students' perspective. For instance, the Students' Evaluation of Educational Quality (SEEQ; Marsch, 1982; Marsh et al., 2009) is a scale composed of 35 statements which address nine aspects of teaching: learning value, enthusiasm, organization, group interaction, individual rapport, breadth of coverage, exams and grading, assignments, workload or difficulty. Toland & De Ayala (2005) developed the Students' Evaluation of Teaching Effectiveness Rating Scale (SETERS), whereas Keeley et al. (2010) created the Teaching Behavior Checklist.

As can be seen, these general instruments are abundant. However, student perspectives in quality of dissertation tutoring requires a distinct level of analysis. For instance, the elaboration process of dissertations requires a more autonomous work by the student, which may affect the connectedness with the role of the professor as a tutor. Moreover, students usually must prepare an oral exposition of the dissertation before a court, which decides the final student qualification. These procedures give dissertation a new perspective, whereby students are supposed to show all the general and specific competences developed in the undergraduate or master's degree.

At the international level, some methodological processes have been raised to dynamically measure the evaluation consistency in faculties or universities (Bettan-Saltikov et al., 2009). However, these processes refer to the criteria analysis without students' perspective. There have also been studies which addresse effective interventions in cognitive or non-cognitive students' revision strategies (Butterfield et al., 1996; Couzijn & Rijlaardsdam, 2005) for the improvement of specific tutoring in writing tasks (Castelló et al., 2011 Maher et al., 2008). Nevertheless, no previous psychometric studies have been found in the field of evaluation of tutoring in undergraduate/master's degree final dissertations beyond these writing tasks. For this reason, the aim of the present study is threefold: (1) to develop a new scale on the students' evaluation of dissertation tutoring, called the Student Evaluation Of Dissertation Tutoring Scale (SEDITUS); (2) to conduct a first pilot study of this new scale in a sample of university students; and (3) to conduct a validation study of the internal structure of the instrument in a large sample of university students.

Our analyses include an assessment of item performance and dimensionality of the scale using item response theory (IRT). Concretely, for the first pilot study, we used the Rasch model (Rasch, 1960), as it provides a method based on the calibration of ordinal data from a shared measurement scale. Moreover, as we used a larger sample with students from different faculties within a university, the second study implements multilevel Rasch modelling. This extended model provides more reliable measures as it considers the nested structure of data, which may affect variances and standard error estimation (Lamprianou, 2013).

STUDY 1

METHOD

Participants

The sample of the first pilot study consisted of 82 university students. From this total, 72% were women; mean age was 24.68 years with standard deviation of 4. All these students were enrolled at the Faculty of Education from the University of Alicante. Concretely, 32.9% were studying the master's degree of Secondary Education, 36.6% the undergraduate degree in Primary Education, 29.3% the undergraduate degree in Child Education, and 1.2% the master's degree in Educational Research.

Measures

Following guidelines for scale development, essential content and format specification were established prior to item development (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 2014; Wetzel & Greiff, 2018). With respect to content specification, item development was guided by the definition of student-teacher evaluation with unique emphasis on dissertation tutoring. To achieve this issue, legal decree and specific guidelines for final undergraduate degrees and master's dissertations from different universities were analysed. The following content was considered as relevant. First, the focus on the professor's tutoring behaviours in quality and content (support tasks); second, professors' correction tasks in appropriate timing (feedback); and third, support in academic procedures (e.g., document format, timings, guidance on the virtual platform).

Regarding format specifications, the final instrument consisted of a Likert scale with four response categories expressing the level of agreement (1 = low agreement; 4 = total agreement). This format fits previous studies evaluating students' perceptions, as items with ordinal responses were used to fit the construct and provide more information (García-Moya et al., 2020). The items of the instrument can be seen in Table 1.

Table 1: Final items of the student evaluation of dissertation tutoring scale (SEDITUS).

	My dissertation tutor…
1	has informed me about the general dissertation guidelines
2	has advised me on the dissertation subject
3	has responded to my e-mails or virtual tutorials
4	has corrected my drafts in timely manner
5	has given to me the possibility of having face-to-face tutoring
6	has informed me of the dissertation qualification criteria
7	has informed me of the procedure to defend the dissertation

Procedure

The instrument was administered online due to COVID-19 restrictions in 2020 (academic year 2019-2020). For this purpose, a Google sheet form was obtained, including the instructions for completion. Participants were informed about the voluntariness of filling in the scale and the confidential use of data by e-mail. A total of two reminder e-mails were sent within a 90-day period.

Data analysis

We used a Rasch model (Rasch, 1960) for the initial analysis of the pool of items. The Rasch model is a probabilistic model based on the calibration of ordinal data from a shared measurement scale which enables one to test conditions such as dimensionality, linearity, and local independence (Wright, 1997). As using the same measurement scale, homogeneous intervals can be obtained for both parameters of item difficulty and subject ability. In the present study, four-point Likert-type items were tested, considering the ordered response alternatives as invariant for all items in the scale. These assumptions can be followed through the Andrich Rating Scale Model (RSM) (Andrich, 1978; Wright & Masters, 1982), a specific Rasch-family IRT model that can handle polytomously scored item response data. To fit the RSM assumption, all items need to have the same number of options or categories, and it assumes that adjacent threshold parameters are equally spaced across all items.

Outfit and infit statistics, as well as Rasch reliability, were used to check the quality of the scale from a Rasch measurement perspective. These indexes are measures of the extent to which the data match specifications of a Rasch model. Mathematically, they are the mean value of the squared residuals. A residual is the difference between a subject's response to a given item and the expected response calculated by the model. Therefore, the larger the squared residual, the larger was the misfit between data and model. The difference between infit and outfit is based on the way they are computed. Infit statistics give more importance to those items which are aligned with the person's ability level, as they can carry more information about the person's ability. Values of outfit and infit mean squares (MNSQ) can range from 0 to positive infinitive. Values below 1 indicate a higher than expected fit to the model, while values greater than 1 indicate a poor fit of the model. The category's function of the rating scale was also examined, according to the monotonic increasement of the thresholds and the count of answers for each category (Linacre, 2002). These analyses were done with the package mirt in R free software, developed by Chalmers

Table 2: Bivariate correlations of the items.

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8
Item 1	-							
Item 2	.61**	-						
Item 3	.42**	.39**	-					
Item 4	.37**	.44**	.48**	-				
Item 5	.40**	.42**	.81**	.37**	-			
Item 6	.57**	.61**	.46**	.42**	.43**	-		
Item 7	.36**	.35**	.37**	.57**	.27*	.55**	-	
Item 8	.52**	.51**	.43**	.39**	.40**	.58**	.52**	-

Table 3: Difficulty parameter, fit indexes and item-test correlation.

Item	Difficulty	Infit	Outfit	Item-test corr
1	-.54	.96	1.04	.65
2	-.11	.95	.91	.69
3	-.63	.76	.67	.67
4	-.78	1.10	1.06	.72
5	-.23	.95	.91	.69
6	.10	.99	.79	.73
7	.71	1.10	1.16	.70
8	-.08	.92	.92	.70

(2018). Lastly, to further investigate the psychometric properties of the scale, differential item functioning (DIF) of items can be analysed. The existence of DIF indicates that different groups may have different interpretation of perspective on the items. This study was used to investigate the extent to which male and female students have performed differently on the same items through an ordinal (common odds-ratio) logistic regression, using IRT theta estimates as the conditioning variables. Items flagged for DIF are treated as unique items and group-specific item parameters are obtained. This iterative process is helpful because it only flags items after two consecutive iterations (it ensures that a flagged item cannot affect the selection of another items). All these analyses were done with the package lordif in R free software (Choi et al., 2011).

RESULTS

In the first place, bivariate correlations are shown in Table 2. All correlations were positive and statistical significance, but less than .85, which is the cut-off criteria to item exclusion. Therefore, all the items were kept for subsequent analysis.

With respect to RSM, dimensionality was assessed through standardized residuals, also called eigenvalues. According to Linacre (2012), values below 2 in the first contrast indicates that residuals are not relevant enough to worsen dimensionality of the instrument. In our scale, we obtained a first contrast of 2.0 eigenvalues, which confirms the unidimensional nature of the instrument.

In the second place, items' fit to the model was checked with Infit and Outfit values (Table 3). According to Bond & Fox (2015), appropriate values range from 0.7 to 1.3 in this type of scale. As can be seen, Infit and Outfit values are within an appropriate range for accepting item fit to the model. On the other hand, difficulty levels spread across the ability continuum from negative to positive logit values. The easiest item was item 1 ("My tutor has informed me about the general dissertation guidelines"), which means that most participants need a low level of the construct of satisfaction to express high values in this item. The most difficult item was item 7 ("My tutor has informed me of the dissertation qualification criteria"), meaning that a high level of satisfaction is needed to obtain higher scores in this item.

With respect to the category's function of the rating scale (Table 4), it was confirmed step calibration increased

Table 4: Summary of category structure.

Category	Observed count	%	Observed average	Infit MNSQ	Outfit MNSQ	Andrich Threshold	Category measure
1	93	14	-1.00	1.18	1.27	-	-2.02
2	102	16	-.12	.95	.81	-.62	-.56
3	146	23	.62	.97	.76	-.04	.55
4	307	47	1.58	.99	.97	.66	2.04

Table 5: Sample distribution per faculty.

Faculty	Men	Women	Total
Faculty of Education	56 (27.86%)	201 (72.14%)	257
Faculty of Philosophy	46 (27.05%)	124 (72.95%)	170
Faculty of Economic Sciences	48 (29.62%)	114 (70.38%)	162
Higher Polytechnic School	82 (62.12%)	50 (37.88%)	132
Faculty of Sciences	38 (28.78%)	55 (71.22%)	93
Faculty of Law	42 (27.45%)	111 (72.55%)	153
Faculty of Health Sciences	12 (15.18%)	67 (84.82%)	79
Total	324 (30.97%)	722 (69.03%)	1046

monotonically to ensure that higher measures on the items represented higher trait under measurement, with threshold values of -.62 between categories 1 and 2; -.04 between categories 2 and 3; and .66 between categories 3 and 4. Infit and Outfit values were excellent for all categories (between 0.7 and 1.3).

The analysis of DIF estimated the distribution of the difficulty parameter in the sample of males and females. The ordinal logistic criterion used a chi-square distribution with an alpha cut-off value of .01. After the first iteration for purification procedure, no flagged items were detected for any item.

STUDY 2

METHOD

Participants

After the initial pilot study, a more extended validation study of the internal structure of the instrument was implemented in a large sample of 1046 students from all the faculties enrolled at the University of [BLINDED]. Table 5 shows the distribution of frequencies and gender percentages per faculty. There is a higher percentage of women (69.03%) in all the faculties with the exception of the Higher Polytechnic School.

Measures

The instrument was re-evaluated by a committee of professors from all the faculties involved. Item quality was assessed in terms of how the content relates to the specific tutoring systems for the competencies developed in all the university degrees to ensure transversal meaning and linking with objectives. Agreement was reached to remove item 8 ("My dissertation tutor has recommended bibliography according to the topic"), as it was considered too ambiguous to determine tutoring quality in many scientific fields. The remaining items were considered appropriate to be applied to all the sample. Therefore, the final instrument was composed of seven items.

Procedure

The study was approved by the protecting data unit of the University of Alicante, including information about the objectives and requesting an encrypted e-mail address list for all students enrolled in the subjects of final dissertations in undergraduate degrees or master's degrees during academic year 2020-2021. After permission, an invitation e-mail was sent with the project objectives and a request to voluntarily participate. This invitation was sent in May, as students were in the final period of their dissertation and could develop a complete view of their respective tutoring processes.

Data analysis

For the subsequent analysis, multilevel Rasch modelling was applied, considering students to be nested in faculties (Lamprianou, 2013). Multilevel IRT models are based on the estimation of random or hierarchical models (Chaimingkol et al., 2007; Fox & Verhagen, 2010). Considering items as polytomic variables, two possible models were proposed: a first, simpler model (Model 1) in which no specific standard deviation per item was computed, and a second model, which includes the estimation of items' standard deviation (Model 2). Thus, the influence of each faculty in students' response pattern can be checked. Deviance Information Criterion (DIC) was used to analyse the most appropriate model. In the second place, item fit Rhat index were used, considering as appropriate those values below 1.1 (Luo & Jiao, 2018). The Expected a Posteriori (EAP) measure was also used, having a similar interpretation as the reliability coefficient. All the analysis was made with the sirt package in R free software (Robitzsch, 2020).

Finally, differences in item performance between men and women were compared via t-test, considering each faculty as independent samples. Differences between faculties were also explored via one-way ANOVA, followed by the post-hoc Games-Howell test, which is appropriate when there are groups with differing numbers of subjects and equal variances are not assumed (Tabachnick & Fidell, 2007).

RESULTS

In the first place, DIC values were compared to determine the best model to implement item calibration. Model 1 had a DIC value of 11126.04, whereas Model 2 had a DIC value of 10185.7. It can be inferred that Model 2 is the most appropriate to the estimation of difficulty parameters, considering response patterns which depend on the faculty where students are enrolled. Moreover, Model 2 had an EAP of 0.862, which is a consistent value for the instrument to be considered as reliable. Table 6 shows low item difficulty parameters of the model, implying all students can easily understand the item content and give adequate values.

Table 6: Difficulty parameter, standard deviation, and Rhat values of Model 2.

Item	*b* parameter	Standard deviation	Rhat
1	-3.16	.02	1.00
2	-3.14	.02	1.00
3	-3.21	.02	1.00
4	-2.99	.02	1.00
5	-2.92	.03	1.00
6	-2.56	.02	1.00
7	-2.54	.02	1.01

After checking the model fit, statistical significance tests were implemented between men and women per item and per faculty. T-tests for independent samples were used by selecting the faculty samples separately. Results indicate that significant differences existed only in item 7 for female students enrolled in the High Polytechnic School (t = 2. 83, df = 130, $p \leq .05$).

Finally, Table 7 includes descriptive statistics, by which a one-way ANOVA was implemented to analyse possible differences between faculties with respect to the total score obtained in the scale. The results indicate that significant differences were only found between the Faculty of Education and the Higher Polytechnic School (mean differences = -2.26, $p \leq .05$); and between the Faculty of Law and the Higher Polytechnic School (mean differences = -2.50, $p \leq .05$).

DISCUSSION

The aim of this research was to use IRT modelling to examine the psychometric properties of a new instrument

Table 7. Means and standard deviations (SD) per item distributed across faculties.

Item	Faculty of Education	Faculty of Philosophy	Faculty of economic sciences	Faculty of Sciences	Faculty of Law	Faculty of Health Sciences	Higher Polytechnic school
1	3.19 (1.00)	3.08(1.04)	3.09(1.98)	3.33(0.85)	3.30(0.85)	3.06(1.10)	3.33(1.00)
2	3.02(1.11)	3.18(1.07)	3.07(1.17)	3.36(0.91)	3.23(0.92)	3.10(1.11)	3.36(0.85)
3	3.07(1.17)	3.36(1.00)	3.12(3.40)	3.40(0.94)	3.38(0.96)	3.11(1.21)	3.40(0.95)
4	2.85(1.23)	3.08(1.14)	2.93(1.24)	3.20(1.06)	3.05(1.08)	2.90(1.27)	3.20(1.07)
5	2.81(1.27)	2.86(1.16)	2.78(1.28)	3.23(1.05)	3.33(1.01)	2.86(1.24)	3.23(1.05)
6	2.48(1.22)	2.54(1.21)	2.56(1.24)	2.89(1.14)	2.56(1.21)	2.43(1.26)	2.89(1.14)
7	2.57(1.24)	2.62(1.24)	2.54(1.23)	2.85(1.13)	2.42(1.16)	2.28(1.24)	2.85(1.13)

developed to measure students' satisfaction of tutoring, based on the use of transversal competences determined at the university system. The two studies represent relevant psychometric investigation and the first IRT modeling effort in the general field of dissertation tutoring. IRT has been extensively used in the field of educational psychology, such as in student evaluation of teaching (Sánchez et al., 2021) or emotional intelligence (Cooper & Petrides, 2010; Rubio et al., 2007), and it can be helpful in developing and evaluating short measures, matching the objective of this study.

In this setting, the authors of the present study pretend to explore the importance of students' perspective when assessing dissertation tutoring, as a new assessment criterion apart from traditional assessment (Bettany-Saltikov et al., 2009). Modern perspectives of student evaluation of teaching could then be applied to dissertation tutoring, considering beliefs, knowledge and actions as main concepts determined by contextual influences (Soraya & Amundsen, 2001). Noting this perspective, in Spanish higher education institutions, special efforts have been made to consider students as relevant agents of social and academic development.

Taken together, the studies suggest the final instrument of seven items has appropriate psychometric properties, considering the nested nature of the data (students enrolled in faculties) with a multilevel IRT model when a more extended sample is used. The functioning of response categories indicated that the four-category structure did function quite well. It was also observed that the model needs to take into account the items' estimated variance, which is the item variability level between faculties. At the same time, fit parameters show acceptable values, indicating the items are measuring a factor of students' level of satisfaction of dissertation tutoring. Evidence of good psychometric properties comes from the gender DIF analysis, and the impact of gender difference was taken into consideration in this study. No substantial gender DIF was found for the remaining seven items of the scale.

The analysis also indicates some gender differences; concretely in item 7 between men and women enrolled in the High Polytechnic School. This difference, however, cannot be generalized to other faculties, as it can be considered a minor difference. Future research may further explore this difference by considering new students' cohorts from different years, together with alternative bias analysis.

Despite the positive outcomes of the study, some limitations may need to be addressed in the future. In the first place, it should be noted that the data used herein were students enrolled only at the University of Alicante. Larger samples from other Spanish universities may enable both better estimates and comparison of the constructs across different social or economic contexts. Secondly, some important socio-demographic or academic variables may impact on the way students value the satisfaction level of dissertation tutoring. For instance, it is possible that teachers with more experience in tutoring (e.g., number of years) give more comprehensive feedback or support materials. For this reason, further analysis is required to detect possible causes of students rating between sample subgroups and efficient vs non-efficient teachers' regarding features associated with dissertation tutoring. Such work would further both the reliability and validity of the current measurement using diverse populations and contexts. Last, it should be necessary to analyze whether previous and current students' achievement levels affect dissertation tutoring based on professor-student relationship. Recent studies have indicated that achievement levels moderate the positive relationship between individualized Teacher Frame of Reference (iTFR) and self-concept (e.g., Helm et al., 2022; Lüdtke et al.,2005). At the dissertation tutoring level, professors' characteristics may be even more important to determining positive students' self-concept (Dietrich et al., 2005). This scale may be considered in further studies to analyze these relations in different time points (beginning and end of dissertation tutoring periods) and provide useful feedback in order to change strategies for intervention.

Aknowledgement

This research has been granted by the Institute of Education of the University of Alicante (Reference of the Grant: 5198).

References

Andrich, D. (1978). A rating information for ordered response categories. *Psychometrica, 43,* 561-573.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing.* American Educational Research Association.

Bettany-Saltikov, J., Kilinc, S., & Stow, K. (2009). Bones, boys, bombs and booze: an exploratory study of the reliability of marking dissertations across disciplines. *Assessment & Evaluation in Higher Education, 34*(6), 621-639. https://doi.org/10.1080/02602930802302196

Boice, R. (1991). New faculty as teachers. *Journal of Higher Education, 62*(2), 150-173.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Routledge.

Butterfield, E., Hacker, J., & Albertson, L. (1996). Environmental, cognitive, and metacognitive influences on text revision: Assessing the evidence. *Educational Psychology Review, 8*(3), 239-297.

Castelló, M., Iñesta, A., Pardo, M., Liesa, E., & Martínez-Fernánez, R. (2012). Tutoring the end-of-studies dissertation: helping psychology students find their academic voice when revising academic texts. *Higher Education, 63*, 97-115. https://doi.org/10.1007/s10734-011-9428-9

Chaimongkol, S., Huffer, F. W., & Kamata, A. (2007). An explanatory differential item functioning (DIF) model by the WinBUGS 1.4. *Songlanakarin Journal of Science and Technology, 29*, 449-458.

Chalmers, P. (2018). *mirt: A multidimensional item response theory package for the R environment*. Retrieved from https://cran.r-project.org/web/packages/mirt/mirt.pdf

Choi, S. W., Gibbons, L. E., Crane, P. K. (2011). lordif: An R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/Item Response Theory and Monte Carlo simulations. *Journal of Statistical Software, 39*(8), 1-30.

Christensen, K. B., Makransky, G., & Horton, M. (2017). Critical values for Yen's Q3: Identification of local dependence in the Rasch model using residual correlations. *Applied Psychological Measurement, 41*(3), 178-194. https://doi.org/10.11770/0146621616677520

Cooper, A., & Petrides, K. V. (2010). A psychometric analysis of the Trait Emotional Intelligence Questionnaire-Short Form (TEIQue-SF) using item response theory. *Journal of Personality Assessment, 92*(5), 449-457. https://doi.org/10.1080/00223891.2010.497426

Couzijn, M., & Rijlaardsdam, G. (2005). Learning to write by reader observation and written feedback. In G. Rijlaarsdam, H. Van den Bergh, & M. Couzijn (eds). *Effective teaching and learning of writing: Current trends in research* (pp. 224-253). Amsterdam University Press.

Dietrich, J., Dicke, A. L., Kracke, B., & Noack, P. (2015). Teacher support and its influence on students' intrinsic value and effort. Dimensional comparison effects across subjects. *Learning and Instruction, 39*, 45-54. https://doi.org/10.1016/j.learninstruc.2015.05.007

Evans, C. Kandiko-Howson, C., Forsythe, A., & Edwards, C. (2020). What constitutes high quality higher education pedagogical research? *Assessment & Evaluation in Higher Education.* https://doi.org/10.1080/02602938.2020.1790500

Fox, J, P., & Verhagen, A. J. (2010). Random item effects modeling for cross-national survey data. In E. Davidov, P. Schmidt, & J. Billiet (eds.), *Cross-cultural analysis: Methods and Applications* (pp. 467-488). Routledge Academic.

García-Moya, I., Brooks, F., & Moreno, C. (2020). A new measure for the assessment of student-teacher connectedness in adolescence. *European Journal of Psychological Assessment, 37*(5), 357-367. https://doi.org/10.1027/1015-5759/a000621

Glaser, R., Lesgold, A., & Lajoie, S. P. (1988). Toward a cognitive theory for the measurement of achievement. In R. Ronning, J. Glover, J. S. Conoley, & J. C. Wittrock (eds). *The influence of cognitive psychology on testing* (vol. 3). Erlbaum.

Helm, F., Wolff, F., Möller, J., Zitzmann, S., Marsh, H., & Dicke, T. (2022). Individualized teacher frame of reference and student self-concept within and between school subjects. *Journal of Educational Psychology*. Advanced online publication. https://doi.org/10.1037/edu0000737

Huybers, T. (2014). Student evaluation of teaching: the use of best-worst scaling. *Assessment & Evaluation in Higher Education, 39*(4), 496-513. https://doi.org/10.1080/02602938.2013.851782

Keeley, J., Furr, R. M., & Buskist, W. (2010). Differentiating psychology students' perceptions of teachers using the Teacher Behavior Checklist. *Teaching of Psychology, 37,* 16-20. https://doi.org/10.1080/00982890342682

Lamprianou, I. (2013). Application of single-level and multilevel Rasch models using the lme4 package. *Journal of Applied Measurement, 14*(1), 1-12.

Larsson, S. (1986). Learning from experience: teachers' conceptions of changes in their professional practice. *Journal of Curriculum Studies, 19*(1), 37-43.

Leinhardt, G., & Greeno, J. G. (1986). The cognitive skill of teaching. *Journal of Educational Psychology, 78*(2), 75-95.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3,* 85-106.

Linse, A. R. (2017). Interpreting and using student rating data: Guidance for faculty serving as administrators and on evaluation committees. *Studies in Educational Evaluation, 54,* 94-106. https://doi.org/10.1016/j.stueduc.2016.12.004

Lüdtke, O., Köller, O., Marsh, H. W., & Trautwein, U. (2005). Teacher frame of reference and the big-fish-little pond effect. *Contemporary Educational Psychology, 30*(3), 263-285. https://doi.org/10.1016/j.cedpsych.2004.10.002

Luo, Y., & Jiao, H. (2018). Using the Stan program for Bayesian Item Response Theory. *Educational and Psychological Measurement, 78*(3), 384-408.

Maher, D., Seaton, L., McMullen, C., Fitzgerald, T., Otsuji, E., & Lee, A. (2008). Becoming and being writers: The experiences of doctoral students in writing groups. *Studies in Continuing Education, 30*(3), 263-275.

Marais, I., & Andrich, D. (2008). Effects of varying magnitude and patterns of response dependence in the unidimensional Rasch model. *Journal of Applied Measurement, 9*(2), 105-124.

Marsh, H. W. (1982). SEEQ: a reliable, valid and useful instrument for collecting students' evaluation of university teaching. *British Journal of Educational Psychology, 52,* 77-95.

Marsh, H. W., Muthén, B., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A. J. S., & Trautwein, U. (2009). Exploratory structural equation modeling, integrating CFA and EFA: Application to students' evaluations of university teaching. *Structural Equation Modeling, 16*, 439-476. https://doi.org/10.1080/10705510903008220

Pajares, M. F. (1992). Teachers' beliefs and educational research: cleaning up a messy construct. *Review of Educational Research*, 307-332.

Pratt, D. D. (1992). Conceptions of teaching. *Adult Education Quarterly, 42*(4), 203-220.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Denmark: Danish Institute for Educational Research (Expanded edition, 1980). University of Chicago Press.

Richardson, J. T. E. (2005). Instruments for obtaining student feedback: a review of the literature. *Assessment & Evaluation in Higher Education, 30*(4), 387-415. http://doi.org/10.1080/02602930500099193

Robitzsch, A. (2021). *sirt: Supplementary item response theory models.* R package version 3.11-21. https://CRAN.R-project.org/web/packages/sirt/sirt.pdf

Rubio, V. J., Aguado, D., Hontangas, P. M., & Hernández, J. M. (2007). Psychometric properties of an emotional adjustment measure: An application of the graded response model. *European Journal of Psychological Assessment, 23*, 39-46.

Sánchez, T., Veas, A., Gilar-Corbí, R., & Castejón, J. L. (2021). Psychometric perspectives in educational and learning capitals: Development and validation of a scale on student eval-

uation of teaching in Higher Education. *Psychological test and Assessment Modeling, 63*(2), 149-167.

Saroyan, A., & Amundsen, C. (2001). Evaluating university teaching: Time to take stock. *Assessment & Evaluation in Higher Education, 26*(4), 341-353. https://doi.org/10.1080/02602930120063493

Stierer, B., & Antoniou, M. (2004). Are there distinctive methodologies for pedagogic research in Higher Education? *Teaching in Higher Education, 9*(3), 275-285. https://doi.org/10.1080/1356251042200021660 6

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics*. Pearson.

Toland, M., & De Ayala, R. J. (2005). A multilevel factor analysis of students' evaluation of teaching. *Educational and Psychological Measurement, 65*, 272-296. https://doi.org/10.1177/0013164404268 66

Webb, M., & Jones, J. (2009). Exploring tensions in developing assessment for learning. *Assessment in Education: Principles, Policy & Practice, 16*(2), 165-184. https://doi.org/10.1080/09695940903075925

Wetzel, E., & Greiff, S. (2018). The world beyond rating scales. Why we should think more carefully about the response format in questionnaires [Editorial]. *European Journal of Psychological Assessment, 34*(1), 1-5. https://doi.org/10.1027/1015-5759/a000469

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis.* MESA Press.

Wyatt-Smith, C., Klenowski, V., & Gunn, S. (2010). The centrality of teachers' judgement practice in assessment: a study of standards in moderation. *Assessment in Education: Principles, Policy & Practice, 17*(1), 59-75. http://doi.org/10.1080/09695940903565610