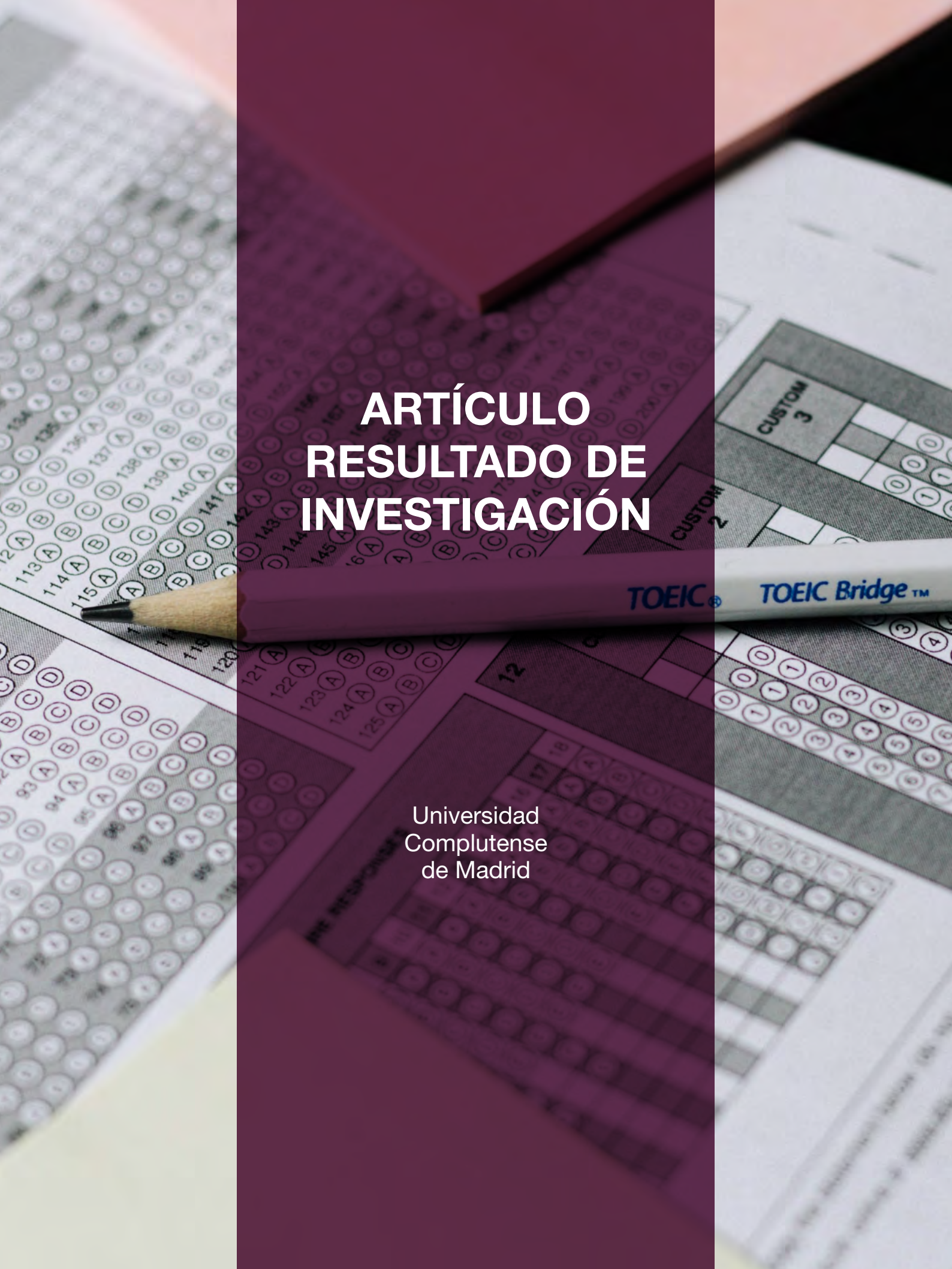

ARTÍCULO RESULTADO DE INVESTIGACIÓN

Universidad
Complutense
de Madrid

Ventajas de aplicar la TRI en cuestionarios sobre la actividad docente universitaria

Advantages of applying IRT in questionnaires on university teaching activity

Vantagens da aplicação do TRI em questionários na atividade docente universitária

Recibido: 03-05-2022 **Aprobado:** 17-05-2022 **DOI:** https://doi.org/10.22490/27452115.5801

AUTORES

Lady Catheryne Lancheros Florián[1]

1. Psicóloga, Universidad Nacional de Colombia. Magíster en Psicología, Universidad Nacional de Colombia. Magíster en Metodología de la investigación para las ciencias del comportamiento y de la salud, Universidad Autónoma de Madrid, España. Candidata a doctora en Neurociencia Cognitiva y Psicología Experimental de la Universidad Complutense de Madrid, España. Docente, investigadora y consultora en Psicometría, Evaluación Educativa y desarrollo de instrumentos de medición psicológica. E-mail: ladycala@ucm.es. ORCID:https://orcid.org/0000-0002-0871-3326

RESUMEN

Los cuestionarios para evaluar la satisfacción de los estudiantes respecto a la labor de los docentes en el ámbito universitario han sido utilizados como herramientas para valorar la calidad de su desempeño y como evidencia de calidad de la enseñanza en los procesos de acreditación institucional. El objetivo del estudio consiste en comparar el análisis psicométrico de un cuestionario sobre la actividad docente universitaria, tipo Likert, utilizando la Teoría Clásica de los Test (TCT) y el Modelo de Respuesta Graduada de Samejima (MRG). Se contó con una base de datos, que contenía las respuestas de los estudiantes de diferentes carreras universitarias pertenecientes a una universidad española. Se realizó el análisis inicialmente usando la TCT y posteriormente con el modelo MRG. Se estimaron y analizaron los parámetros de los ítems con ambos modelos y se resaltó la información que se obtenía con cada uno. En conclusión, esta aplicación permitió identificar que el cuestionario cuenta con adecuadas propiedades psicométricas que permite, con un alto nivel de confiabilidad y discriminación, valorar la satisfacción de los estudiantes hacia sus docentes. Al final se presentan las bondades del uso del MRG para este tipo de instrumentos, así como la información adicional y detallada que proporciona su uso en el contexto de la evaluación educativa.

ABSTRACT

Questionnaires to assess student satisfaction with the work of university teachers have been used as tools to evaluate the quality of their performance and as evidence of teaching quality in institutional accreditation processes. The aim of the study is to compare the psychometric analysis of a Likert-type questionnaire on university teaching activity, using the Classical Test Theory (CTT) and Samejima's Graded Response Model (GRM). A database was available, containing the responses of students from different university courses belonging to a Spanish university. The analysis was performed initially using the TCT and subsequently with the MRG model. The parameters of the items were estimated and analyzed with both models and the information obtained with each one was highlighted. In conclusion, this application allowed identifying that the questionnaire has adequate psychometric properties that allow, with a high level of reliability and discrimination, to assess student satisfaction with teachers. Finally, the benefits of using the MRG for this type of instrument are presented, as well as the additional and detailed information provided by its use in the context of educational evaluation.

RESUMO

Questionários para avaliar a satisfação dos alunos com o trabalho dos professores no ambiente universitário têm sido utilizados como ferramentas para avaliar a qualidade de seu desempenho e como evidência da qualidade do ensino em processos de acreditação institucional. O objetivo do estudo é comparar a análise psicométrica de um questionário do tipo Likert sobre a atividade docente universitária, utilizando a Teoria Clássica do Teste (TCT) e o Modelo de Resposta Graduada de Samejima (MRG). Foi usado um banco de dados, que continha as respostas de estudantes de diferentes carreiras universitárias pertencentes a uma universidade espanhola. A análise foi realizada inicialmente com o TCT e posteriormente com o modelo MRG. Os parâmetros dos itens foram estimados e analisados com ambos os modelos e as informações obtidas com cada um deles foram destacadas. Em conclusão, esta aplicação permitiu-nos identificar que o questionário possui propriedades psicométricas adequadas que permitem, com elevado nível de fiabilidade e discriminação, avaliar a satisfação dos alunos com os seus professores. Ao final, são apresentados os benefícios da utilização do GRM para este tipo de instrumento, bem como as informações adicionais e detalhadas proporcionadas pelo seu uso no contexto da avaliação educaciAonal.

PALABRAS CLAVE:
evaluación docente, Teoría Clásica de los Test (TCT), Teoría de Respuesta al Ítem (TRI), Modelo de Respuesta Graduada de Samejima (MRG), escalas Likert

KEYWORDS:
Teacher evaluation, Classical Test Theory (CTT), Item Response Theory (IRT), Samejima Graduated Response Model (MRG), Likert scales

PALAVRAS CHAVE:
Avaliação do professor, Teoria Clássica do Teste (TCT), Teoria da Resposta ao Item (TRI), Modelo de Resposta Graduada Samejima (MRG), escalas Likert

INTRODUCCIÓN

La evaluación del desempeño de los profesores tiene como objetivo asegurar la calidad de la educación y ser parte de un proceso de mejora continua (Stake, García y Peréz, 2017; Bolívar, 2008). Los cuestionarios que se realizan sobre la actividad docente, que también se conocen como *Student Evaluations of Teaching* (SET), se presentan como una forma estructurada de recolectar retroalimentación de los estudiantes y que evalúan bajo diferentes criterios la forma en que es impartida la docencia en un curso específico; estas evaluaciones son de carácter obligatorio en las instituciones de educación superior y están reglados por la ley (Oermann, Conklin, Rushton, y Bush, 2018). Estos cuestionarios se utilizan como un elemento de evaluación formativa y a la vez sumativa; formativa dado que se obtienen valoraciones y comentarios de los estudiantes sobre los cuales los docentes pueden reflexionar para mejorar el aprendizaje de los estudiantes y también se usan de forma sumativa como medida de éxito que permite a los directivos identificar buenas prácticas de enseñanza, así como aspectos o docentes que requieren algún tipo de apoyo (Holland, 2019).

Pese a su importancia, estos cuestionarios han generado diversas controversias por sus limitaciones de tipo teórico y práctico, referidas a las múltiples finalidades que las instituciones les destinan, la falta de unidad acerca de un modelo de profesor ideal y las dificultades relacionadas con la aplicación que se evidencian con la baja tasa de respuesta entre los estudiantes (Turull y Buxarrais, 2018; Tejedor y Jornet, 2008). Esto ha generado una gran cantidad y variedad de cuestionarios referidos a la evaluación de la enseñanza (Spooren, Brockx y Mortelmans, 2013), así como estudios que debaten acerca de las características personales del docente, la validez, el sesgo y la dimensionalidad de estos instrumentos (Turull y Buxarrais, 2018).

La información sobre la calidad docente se recolecta principalmente a través de estos cuestionarios de opinión, cuyos datos suelen ser utilizados con otros fines adicionales a la mejora del proceso enseñanza y aprendizaje, como son, la acreditación de los docentes, la acreditación de las titulaciones, el insumo para planes de mejora, entre otros (Denson, Loveday y Dalton, 2010; Espinosa, Fernández Sánchez, Sabater, Valdés y García-Fernández, 2017).

Dada la importancia que tienen estos cuestionarios y las decisiones que se toman con ellos, es importante hacer reflexiones sobre su validez, el análisis de sus datos y el alcance de sus interpretaciones. Diferentes investigaciones han identificado las variables que afectan estas evaluaciones como lo son la simpatía que siente el estudiante hacia el profesor (Feistauer y Richter, 2016), la tendencia del estudiante a responder favorablemente o con aquiescencia (Spooren, Mortelmans y Thijssen, 2012), el género y la disciplina del estudiante, entre otros (Boring, Ottoboni y Stark, 2016), si bien existe un gran número de investigaciones sobre las variables que pueden afectar este fenómeno, también se evidencia que a pesar de las pautas para recopilar e interpretar los datos de SET, muchos usuarios de estas evaluaciones no tienen capacitación para el manejo de estos datos, ni conocen sobre la investigación que se realiza (Penny, 2003).

Los análisis estadísticos realizados y los reportes a los docentes de los SET se suelen realizar haciendo uso de la Teoría Clásica de los Test (TCT), este es el modelo por defecto utilizado en la evaluación de la calidad de los *test* y a pesar de la aparición de otros modelos continúa siendo el de mayor uso; sin embargo, presenta algunas limitaciones en relación con la dependencia de los índices que ofrece en función de la muestra que se use en el análisis. Es decir, si la muestra cambia, la información sobre la calidad del instrumento también cambia; esta situación afecta la interpretación de los datos, especialmente cuando las muestras carecen de homogeneidad (Martínez, 2014). En este sentido se considera que las propiedades psicométricas de un *test* pueden tener un estudio más detallado al utilizar otras técnicas, tales como la Teoría de Respuesta al Ítem (TRI) (Abad, Olea, Ponsoda y García, 2011), en el que ha prevalecido el Modelo de Respuesta Graduada (MRG) (Samejima, 2010) para el análisis de ítems polítomos o politómicos, como son las escalas tipo Likert.

A pesar de que el uso, el análisis y la interpretación de estos datos tienen consecuencias para la carrera de los docentes y la mejora de la enseñanza, existe muy poca investigación sobre cómo afecta que los encargados de analizar los datos tengan poco conocimiento estadístico y psicométrico o que realicen inferencias basados en datos descriptivos, por lo que siguen prefiriendo medidas agregadas y generales de la satisfacción de los estudiantes, desconociendo otras herramientas que pueden dar cuenta del fenómeno o de la calidad de los instrumentos que se utilizan para recopilar los datos (Spooren, Brockx, y Mortelmans, 2013). En esta línea, el presente estudio tiene como objetivo evidenciar los aportes que realizan dos modelos de análisis: la Teoría Clásica de los Test y la Teoría de Respuesta al Ítem, y así observar la información que es posible obtener de estos cuestionarios haciendo uso de cada método.

Teoría Clásica de los Test

La TCT asume entre sus supuestos que la puntuación de una persona en un *test* (X), está compuesta por la puntuación verdadera (V) y el error de medida (e). Las deducciones que se hacen de este modelo permiten llegar a fórmulas que estiman las propiedades de

los ítems y los *test* (Muñiz, 2010). El análisis global del *test* implica revisar elementos de la fiabilidad y las evidencias de validez, pero también cuando se quieren analizar los ítems que componen un *test* es necesario conocer si los parámetros de los ítems guardan relación con los parámetros del *test* y para ellos se observan sus índices de dificultad, discriminación y flujo de opciones. La *dificultad* se refiere a la proporción de personas que aciertan un ítem, del total de personas que lo han intentado resolver. La *discriminación* se centra en el poder para diferenciar a las personas que tienen o no la característica medida y de forma más precisa es la correlación entre las puntuaciones de las personas en el ítem y sus puntuaciones en el *test*. Finalmente, el *flujo de opciones*, es decir la proporción de personas que responde cada alternativa de respuesta suele brindar información sobre la distribución de las respuestas de las personas a cada opción, indicador de calidad de la elaboración del ítem (Muñiz, 2018).

La aplicación de la TCT funciona adecuadamente con la mayoría de los datos empíricos, situación que puede ser tanto una fortaleza como una debilidad; dado que es una fortaleza su simplicidad y a la vez una limitación al no poder profundizar con detalle en las conclusiones que permite extraer de los análisis (Muñiz, 2010; Abal, Auné y Attorresi, 2014). Algunas de las críticas más relevantes de la TCT coinciden en que las mediciones de los constructos no son invariantes de los instrumentos, es decir, que diferentes instrumentos diseñados para medir el mismo constructo no son comparables entre sí y además las propiedades psicométricas que se obtienen de un *test*, como la dificultad, discriminación, fiabilidad y validez dependen de las personas con las que se estimen; quiere decir que, si la muestra cambia, las propiedades del *test* también. (Muñiz, 2010).

A pesar de estas limitaciones, la TCT sigue vigente y en algunos casos se utiliza de forma complementaria con la Teoría de Respuesta al Ítem (TRI), para hacer un análisis más exhaustivo de la calidad del *test*. Además, diferentes autores han encontrado correspondencias entre los indicadores de la TCT y sus equivalentes en la TRI (Barbero, Prieto, Suárez y San Luis, 2001; Kramp, 2006).

Teoría de Respuesta al Ítem

Desde la perspectiva de la TRI se revisa qué ítems se ajustan al modelo y si las opciones de respuesta planteada logran diferenciar entre niveles de habilidad. De esta forma se puede indicar si un instrumento cuenta con adecuadas características psicométricas que determinen la calidad del mismo como vía de evaluación válida de un atributo.

Entre los modelos de la TRI, el MRG pretende establecer la localización de cada umbral en el continuo del rasgo latente (Samejima, 2010; Attorresi, et al., 2011). Existen tres elementos importantes en el modelo de Samejima. El primer elemento corresponde a las CCO (Curvas Características Operantes) que representan la probabilidad de elegir una categoría igual o superior a k, que aumenta con el nivel de rasgo. Estas CCO son un paso intermedio para obtener las CCR (Curva de la Categoría de Respuesta) que indican la probabilidad de escoger cada opción k en cada nivel de rasgo. El segundo elemento corresponde con los parámetros de posición, el parámetro que indica la relación entre el nivel de θ y la categoría que tiene la máxima probabilidad de ser elegida, la media de dos parámetros b_k consecutivos indica el nivel de rasgo en el que la probabilidad de elegir la opción k es máxima. (Abad, Ponsoda y Revuelta, 2006) y el parámetro de discriminación (*a*) que señala la relación entre el ítem y el rasgo medido, se denomina también cómo la discriminación y sus valores se suelen encontrar entre 0.3 y 2.5. El tercer elemento corresponde a las medidas locales de precisión, que se evidencian a través de la Función de Información tanto del ítem como del *test* (Abad, Olea, Ponsoda y García, 2011).

Sin embargo, para el uso de la TRI se deben considerar las dificultades respecto a las especificaciones del tamaño de la muestra, la comprensión de los resultados (Asún y Zúñiga, 2008), además del cumplimiento de supuestos o tener ítems que midan apropiadamente en toda la escala del constructo (Abal, Lozzia, Aguerri, Galibert y Attorresi, 2010).

Puntos en común

Uno de los principales aspectos a evaluar cuando se analizan las propiedades de un *test*, es su confiabilidad, que se refiere principalmente, a la precisión de la medida que se realiza con un instrumento y se reporta mediante un indicador llamado coeficiente de fiabilidad. De los coeficientes existentes, el más utilizado en ciencias sociales es el alfa de Cronbach (Zumbo, Gadermann y Zeisser, 2007) el cual brinda información acerca de la consistencia interna del *test*. La investigación ha mostrado que este coeficiente asume tau-equivalencia y le afectan factores como la longitud de la prueba, la covarianza entre los ítems y la varianza de la prueba, así como, que suele aumentar su valor, si la muestra o el número de ítems aumenta (Cortina, 1993; Abad, Olea, Ponsoda y García, 2011). A su vez, existen modelos de fiabilidad basados en el análisis factorial como el coeficiente Omega, en el que para datos unidimensionales se redefine como la proporción de varianza del *test* que explica el factor común (Abad, Olea, Ponsoda y García, 2011), como ventajas se reporta que no asume tau-equivalencia, trabaja con las cargas factoriales y no depende del número de ítems (Ventura-León y Caycho-Rodríguez, 2017).

El coeficiente Omega reporta mejores indicadores de la fiabilidad cuando se trabaja con escalas tipo Likert (Elosua y Zumbo, 2008). En la TRI, también se realiza un análisis de la fiabilidad mediante la función de información, que permite observar qué tan informativo o preciso es el instrumento en los diferentes niveles de rasgo que se miden (Muñiz,2010).

Con el objetivo de presentar los anteriores elementos se realiza el análisis de un cuestionario acerca de la actividad docente, desde las dos perspectivas: la TCT y la TRI, con el fin de observar el funcionamiento global del cuestionario y de los ítems bajo cada escenario, rescatar las bondades e información que brinda cada una y los aportes que realizan a la valoración de la calidad de este tipo de cuestionarios.

MÉTODO

Enfoque y tipo de investigación

Este artículo se sustenta en un enfoque de investigación cuantitativo, de carácter no experimental, clasificado dentro de los estudios instrumentales (Montero y León, 2002), referidos a los estudios encaminados al desarrollo de pruebas y aparatos, dado que se encuentra centrado en el análisis de la información psicométrica que es posible obtener sobre un instrumento.

Base de datos

La base de datos utilizada en esta investigación corresponde a los datos obtenidos de la aplicación del cuestionario que realizaron los estudiantes de una Universidad de Madrid, España, que facilitó la información de forma anónima, a partir de un trabajo de investigación de fin de máster. La universidad ha solicitado mantener el anonimato en las publicaciones respecto a los datos concedidos; se incluyeron ocho facultades y se consolidaron las respuestas de los estudiantes que presentaron el cuestionario en la universidad durante los años 2012 - 2017. Se analizaron 16.950 respuestas de estudiantes, 61,93% fueron mujeres y 38,07% fueron hombres, se tuvo en cuenta un único cuestionario por estudiante y sin valores perdidos.

INSTRUMENTO

El cuestionario satisfacción de la actividad docente de la que proceden los datos analizados en esta investigación está constituido por siete afirmaciones de respuesta graduada tipo Likert y se debe evaluar el nivel de acuerdo o desacuerdo frente a cada una de estas, en una escala establecida desde *Totalmente en desacuerdo* hasta *Totalmente de acuerdo*, e incluyendo un No Procede. La calificación es ordinal de 1 a 5, en donde la máxima puntuación (5) refleja mayor satisfacción con el criterio de desempeño docente. Adicionalmente, se realizan dos preguntas abiertas al final del instrumento. Esta versión del instrumento fue aplicada a los estudiantes durante los periodos 2012 - 2017.

El instrumento diferencia entre satisfacción alta, media y baja hacia la actividad del docente de un curso. Los estudiantes que están en el rango inferior de puntajes se consideran con baja satisfacción hacia la labor del docente. En contraposición, los estudiantes con mayor puntuación reflejan una alta satisfacción por las actividades realizadas por el docente durante el semestre.

Procedimiento

El análisis psicométrico se divide en dos momentos: en el primero se realiza el análisis utilizando la Teoría Clásica de los Test (TCT) y en el segundo se realiza utilizando el modelo de respuesta graduada de Samejima como modelo de la TRI. Los análisis estadísticos y psicométricos se realizaron por medio del lenguaje de programación R versión 3.5.2 (R Core Team, 2017).

Los indicadores que se estimaron, a partir de la TCT, incluyeron el porcentaje de respuestas por opción en cada ítem, la dificultad del ítem con la media de cada uno, la discriminación del ítem a partir de la correlación ítem - resto del *test* (considerando el criterio superior a 0.2 para indicar que el ítem discrimina) y la fiabilidad si se elimina el ítem. Se realizó también la evaluación del índice de fiabilidad, por medio del coeficiente alfa de Cronbach (Abad, Olea, Ponsoda y García, 2011).

Antes de iniciar con el procedimiento de TRI se realizó la evaluación de la pertinencia de aplicar los métodos de la TRI, mediante la comprobación de los supuestos que señala esta teoría. En primer lugar, se cumplió con el criterio del tamaño de la muestra que sugiere que la cantidad de personas sea superior a mil, siendo en este caso 16.950 estudiantes. Además, se realiza un análisis paralelo para confirmar el cumplimiento del supuesto de unidimensionalidad.

La estimación de parámetros con el Modelo de Respuesta Graduada de Samejima (MRG) (R:ltm[gmr], Dimitris, 2006) consistió en que para cada ítem se estimó un parámetro de discriminación (a), y cuatro parámetros de localización (b1, b2 y b3), se construyeron las CCR y la Función de Información para cada ítem, así como la Función de Información para el Test.

RESULTADOS

Teoría clásica de los Test

En la tabla 1 (ver página 64) se pueden observar los indicadores correspondientes al análisis de ítems para cada uno de los que conforman el cuestionario de actividad docente. Se presenta para cada ítem: la puntuación media (dificultad), las distribuciones en los diferentes niveles de respuesta, el nivel de discriminación (Correlación corregida ítem – *test*) y el coeficiente de fiabilidad (Alfa de Cronbach) si se elimina el ítem.

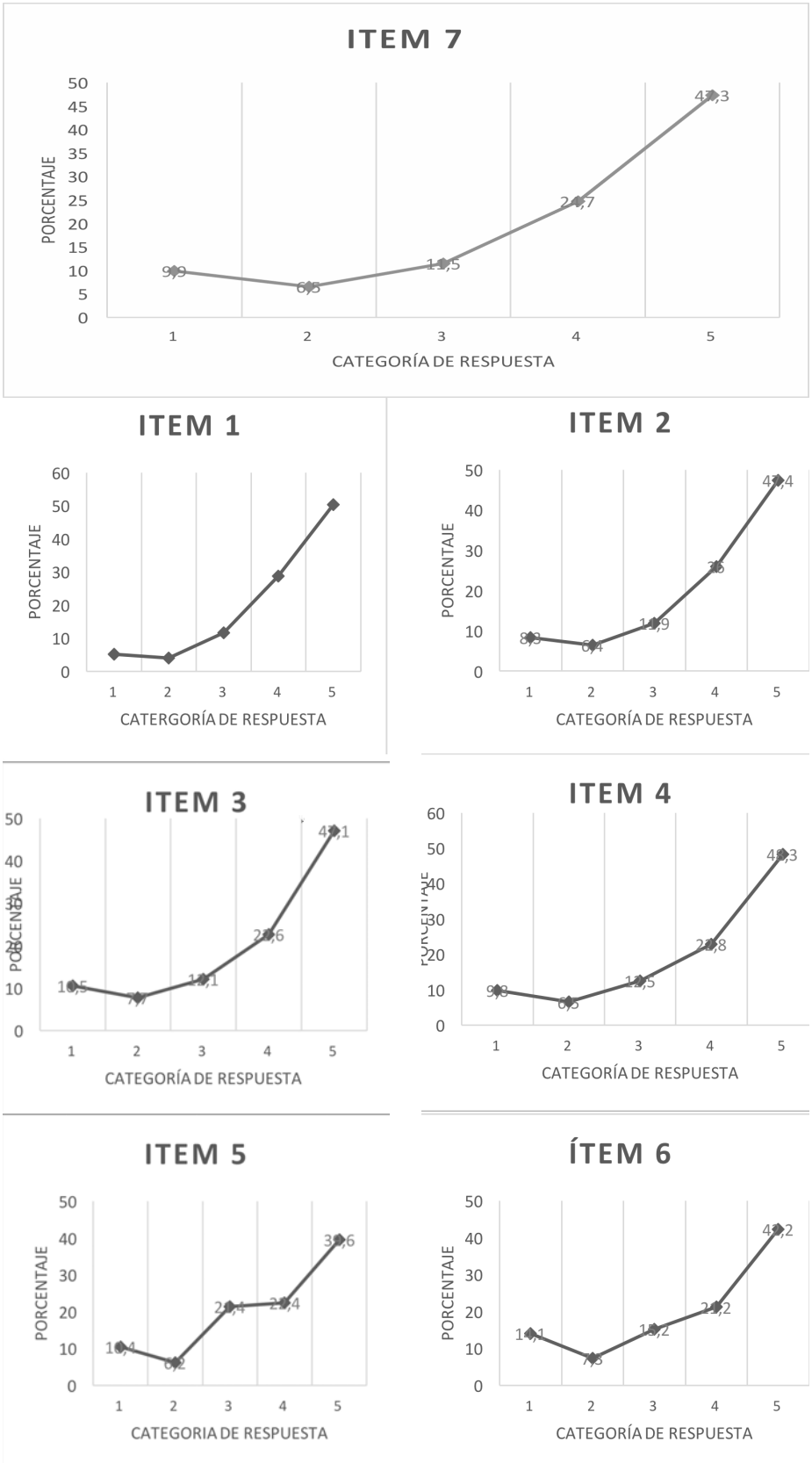

Figura 1. Porcentaje de respuesta de los ítems

Fuente: elaboración propia

Se evidencia que en general todos los ítems presentan una media cercana al valor 4 de la escala ("Más bien de acuerdo"). El *Ítem 1* presenta la media más alta, mientras que el *Ítem 6* la más baja, el cual también evidencia una mayor desviación típica. La correlación ítem - resto del *test* es positiva y con valores superiores a 0.80 en todos los ítems, siendo el *ítem 7* el que evidencia mayor poder discriminativo. No se evidencia que con la eliminación de alguno de los ítems se mejore el coeficiente alfa, ya que el valor global del cuestionario es de $\alpha = 0.96$.

Elección categoría de respuesta

En relación con el porcentaje de respuesta, entre el 40% y el 50% de los estudiantes selecciona la categoría 5, seguido de la categoría 4, tal como se puede ejemplificar con el *ítem 7*, el 47,3% de los estudiantes seleccionaron la categoría 5; el 21,2% la categoría 4; el 15,12% la categoría 3; el 7,3% la categoría 2 y el 14,1% la categoría 1, tal como puede observarse en la figura 1.

Al observar los seis ítems restantes se evidencia un comportamiento similar, en el que en todos los ítems la categoría de respuesta más elegida es la cinco (5); es decir, *Totalmente de acuerdo*, frente al criterio de satisfacción en cada ítem, tal como es posible observar en la figura 1.

Con el fin de determinar la estructura factorial del cuestionario de actividad docente se realizó el análisis paralelo, propuesto por Horn en 1965 (*R:psych*, Revelle, 2017) y se realizó la estimación por el método componentes principales.

Tabla 1.
Descriptivos del análisis de ítems utilizando la TCT

ITEM	Media	Desv. Típica	Correlación Corregida Ítem-resto del *test*	Alfa si se elimina el ítem
1	**4.15**	1.11	0.8	0.96
2	3.98	1.26	0.86	0.96
3	3.88	1.35	0.88	0.96
4	3.93	1.32	0.87	0.96
5	3.75	1.32	0.82	0.96
6	**3.7**	1.43	0.88	0.96
7	3.93	1.32	**0.93**	0.96

Fuente: elaboración propia

Tabla 2.
Autovalores de análisis paralelo

	Autovalores de la Muestra	Media de los valores aleatorios	Percentil 95 de autovalores aleatorios
C1	5,635	1,028	1,034
C2	0,422	1,018	1,024
C3	0,299	1,008	1,015
C4	0,201	1,000	1,006
C5	0,186	0,992	0,997
C6	0,159	0,982	0,995
C7	0,090	0,971	0,979

Fuente: elaboración propia

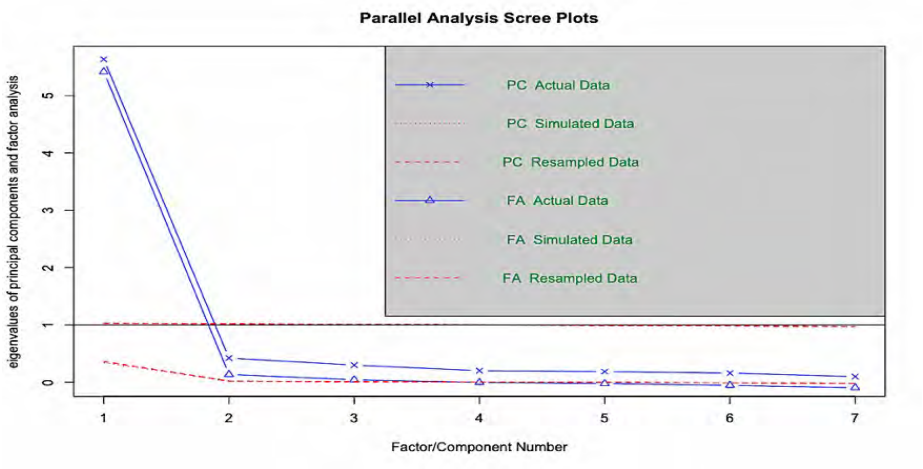

Figura 2. Gráfico análisis paralelo
Fuente: elaboración propia

Los resultados de la tabla 2, en conjunto con los gráficos de la figura 2, sugieren que por el método del análisis paralelo es recomendable retener un único factor, tanto al comparar con el promedio de las matrices aleatorias, como con el percentil 95. Se observa que el único autovalor empírico superior al obtenido aleatoriamente es el primer autovalor (5,635), lo cual también se evidencia en el gráfico, en el que se representa el cambio de pendiente en el paso del autovalor 1 al autovalor 2, ya que a partir del segundo se estabilizan las cantidades de los autovalores. El análisis paralelo señala como mejor solución una estructura unidimensional para el cuestionario de actividad docente.

TEORÍA DE RESPUESTA AL ÍTEM: MRG

Estimación de parámetros

Luego de confirmar la estructura dimensional del *test*, se estimaron los parámetros para el Modelo de Respuesta Graduada (MRG) acorde con el ajuste a este modelo. Los ítems del cuestionario contenían cinco categorías de respuesta por lo que el parámetro b1 se entiende como el mínimo valor del nivel de rasgo necesario para tener una probabilidad mayor de 0.5 de responder "Totalmente en desacuerdo", así que ese valor es el umbral que separa las categorías "Totalmente en desacuerdo" y "Más bien en desacuerdo". De la misma forma b2 separa las categorías en "Más bien en desacuerdo" y "Ni de acuerdo ni en desacuerdo"; b3 diferencia las categorías "Ni de acuerdo ni en desacuerdo" y "Más bien de acuerdo"; y b4 las categorías "Más bien de acuerdo" y "Totalmente de acuerdo". La estimación de los parámetros se realizó con máxima verosimilitud y se observan en la tabla 3.

Tabla 3.
Estimación de parámetros con el MRG

	Estimación parámetros de los ítems				
ITEM	b1	b2	b3	b4	A
1	-2.63	-1.72	-0.82	0.127	3.025
2	-1.69	-1.16	-0.584	0.273	3.348
3	-1.44	-0.94	-0.423	0.279	3.256
4	-1.54	-1.059	-0.498	0.243	3.145
5	-1.59	-1.121	-0.216	0.505	2.582
6	-1.27	-0.801	-0.2	0.409	3.446
7	-1.46	-0.965	-0.486	0.267	4.526

Fuente: elaboración propia

En los parámetros de los ítems se identifica que el parámetro de discriminación (*a*), tiene los valores más altos para el ítem 7, que el ítem denominado de satisfacción global "*En general, el trabajo llevado a cabo por el/la profesor/a ha sido satisfactorio*" esto es posible de evidenciar en la Curva Característica de Respuesta (CCR), en la que cada línea corresponde a la probabilidad de responder a una de las cinco categorías de respuesta en función del nivel de θ. (ver Figura 3). En todos los ítems, el parámetro de discriminación (*a*), es superior al criterio usual utilizado (a>2.5).

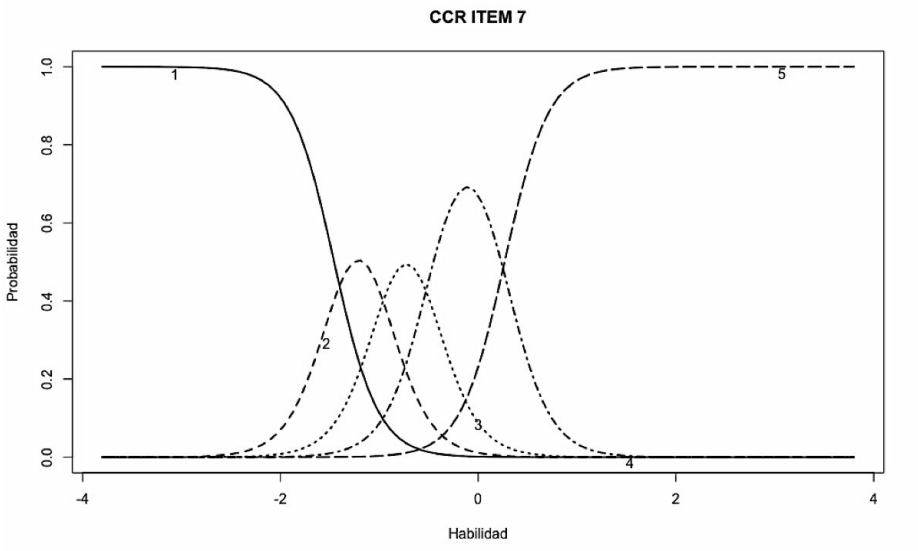

Figura 3. Curva característica de respuesta, ítem 7
Fuente: elaboración propia

Curva Característica de Respuesta (CCR)

En el análisis de las CCR, la línea que se encuentra más a la izquierda corresponde a la probabilidad de responder la categoría que implica menor nivel de rasgo, así mismo la función de la máxima categoría de respuesta tiene mayor probabilidad de ser elegida, entre mayor sea el nivel de rasgo. Se evidencia que el ítem 1 "*El/La profesor/a ha cumplido con lo explicitado en la guía docente*" es más difícil encontrar estudiantes que señalen la categoría "Totalmente en desacuerdo", dado que se requiere muy poco nivel de rasgo para puntuar bajo en ese ítem. Por otra parte, en el ítem 5 "*Las tutorías académicas con este/a profesor/a han resultado útiles*" es el ítem más difícil puntuar en la categoría "Totalmente de acuerdo" dado que solo las personas con un nivel de rasgo muy alto de satisfacción eligen la máxima categoría. (Ver Figura 4).

Al observar los parámetros de los ítems se evidencia que no se requiere un alto nivel de satisfacción para elegir la máxima categoría; es decir, brindar una calificación de 5 a los ítems. Esta información se puede detallar en la figura 11 en el análisis de las CCR del ítem 1 (más fácil) y 5 (más difícil) respectivamente.

Función de información del ítem y del test

Las gráficas de precisión están principalmente representadas en dos elementos: la Función de Información de los Ítems (Figura 5) que como su nombre indica señala la precisión con la que el ítem mide el rasgo latente a lo largo de todo el rango de valores; mientras que la Función de Información del Test (Figura 6), es la sumatoria de las diferentes funciones de información del ítem y muestra en general la precisión de la escala total para discriminar entre los niveles de rasgo de los individuos (Abal, et al., 2014).

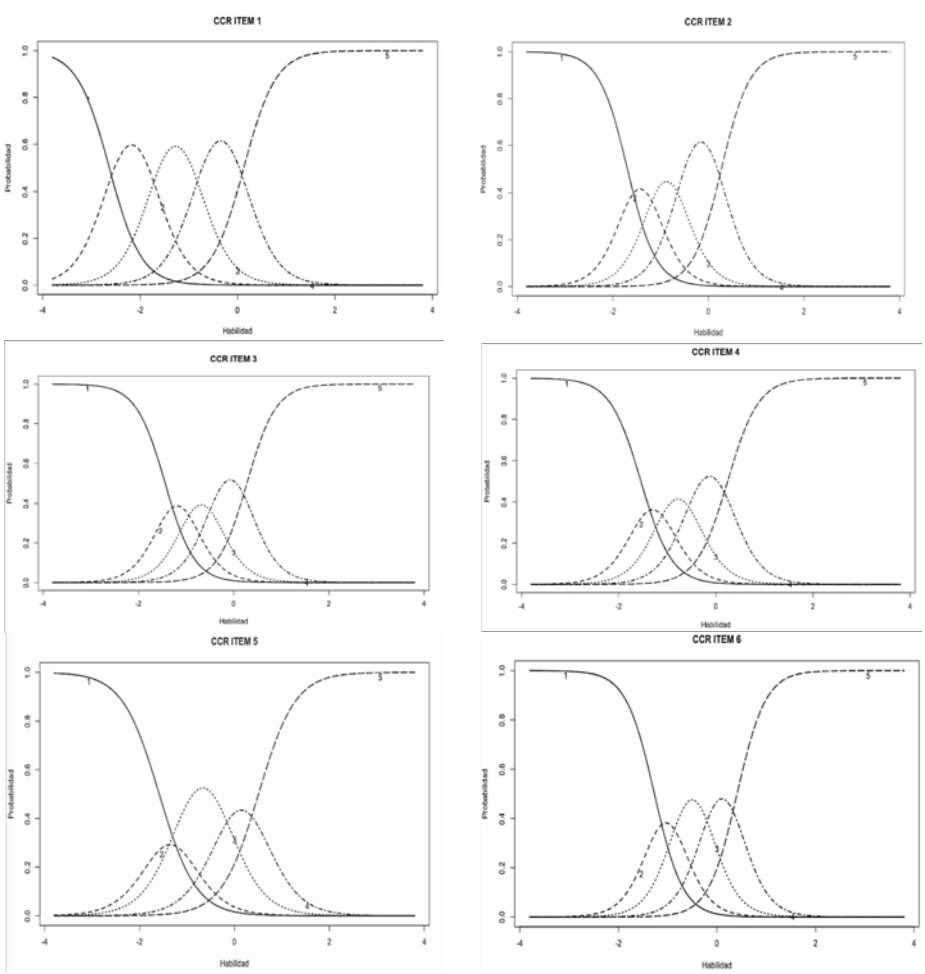

Figura 4. Curva característica de respuesta para 6 ítems.

Fuente: elaboración propia

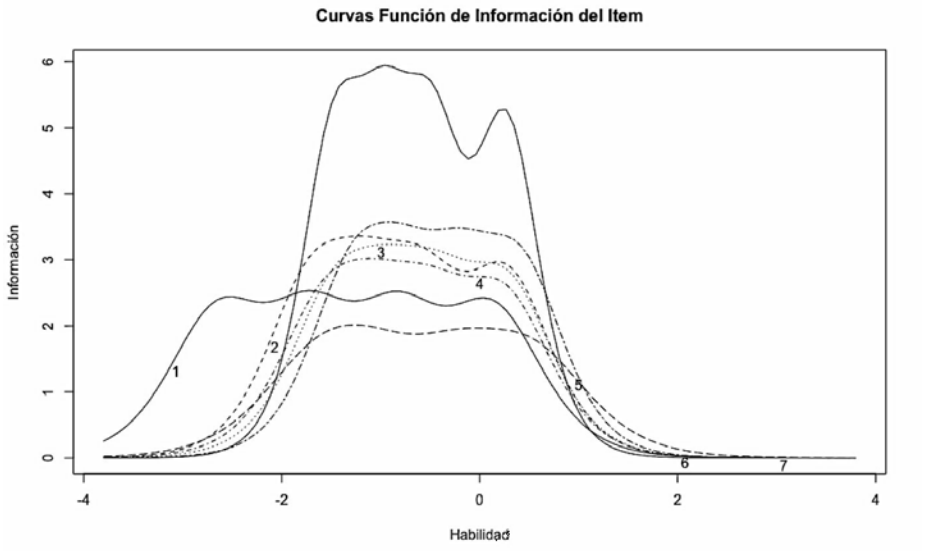

Figura 5. Función de Información de los 7 Ítems

Fuente: elaboración propia

En la figura 5 es posible observar la función de información para todos los ítems; la mayoría miden mejor en los niveles de rasgo medios y bajos, el ítem que más información brinda sobre la satisfacción y para todos los niveles de rasgo es el 7, los que menos información aportan son el 5 y el 1, evidenciando el mismo nivel de información en todo el continuo del rasgo. También se evidencia que el ítem 1 brinda más información para los niveles bajos de satisfacción.

En la figura 6 de la Función de Información del w se observa que en general este proporciona un alto nivel de información, en especial entre -1.5 y 0.5.

Se evidencia a partir de la Función de Información del Test, que este cuestionario de actividad docente proporciona mayor información en los niveles de rasgo medio - bajo.

DISCUSIÓN Y CONCLUSIONES

La evaluación de la calidad del instrumento desde las dos perspectivas TCT y TRI, permitió encontrar adecuados indicadores que convergen en cuanto a evidenciar la calidad del instrumento y que proporcionan evidencias de validez y fiabilidad del cuestionario de evaluación docente y de las interpretaciones que se pueden realizar de sus puntajes en cada uno de los ítems que la componen.

Desde la perspectiva clásica (TCT), el análisis de los ítems muestra que estos tienen una media en torno al valor 4 de la escala, lo que sugiere que los estudiantes están “Más bien de acuerdo” con las afirmaciones del cuestionario de actividad docente y que el mayor porcentaje de las respuestas se encuentra en la categoría 5 “Totalmente de acuerdo”, también se evidenció un alto nivel de correlación entre cada uno de los ítems y el resto del *test* (>0.8). En términos generales estos indicadores reflejan un alto nivel de satisfacción con labor del docente.

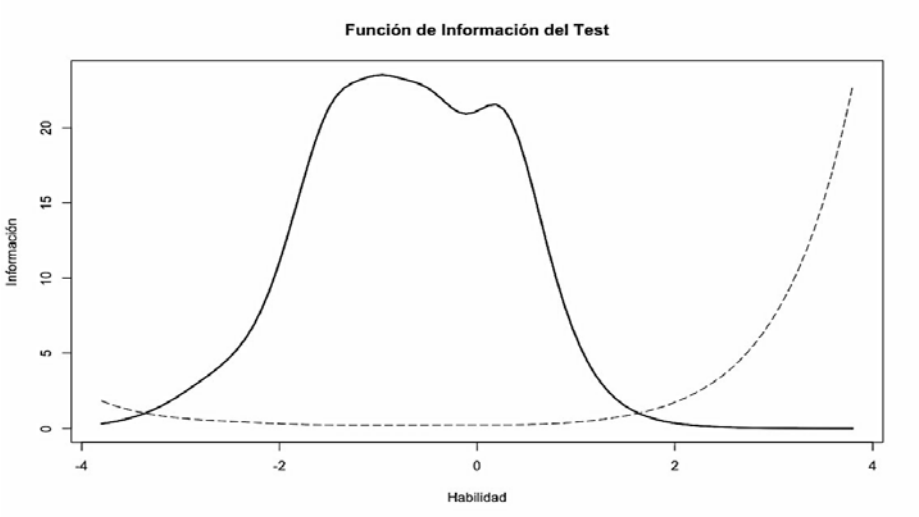

Figura 6. Función de Información de la escala completa de 7 Ítems
Fuente: elaboración propia

Por su parte, el análisis paralelo y los pesos factoriales superiores a 0.8, dan muestra de una estructura unidimensional, siempre es recomendable apoyarse en diversas fuentes y compararlas para tener evidencia de la decisión sobre la estructura factorial, ya que se conoce, que los criterios de ajuste estadístico no siempre apoyan las mejores decisiones (Abad, Olea, Ponsoda y García, 2011), por ello una buena forma de complementar los análisis fue observar el análisis paralelo y los datos referentes a los autovalores.

El análisis realizado desde la TRI con el MRG ofrece otros indicadores sobre la calidad del instrumento. Además de identificar el nivel de atributo necesario para seleccionar una categoría, que se considera equivalente al nivel de dificultad de la TCT en el que el ítem más fácil es el 1 y el más difícil es el 5; este modelo proporciona el indicador de discriminación que en este caso concreto es superior para todos los ítems (>2.5). Adicionalmente, permite identificar los ítems más informativos que, para este caso, fueron los ítems 7, 6, 3, 2 y 4, mientras que aquellos que aportan menos información sobre la satisfacción de los estudiantes son los ítems 1 "*El/La profesor/a ha cumplido con lo explicitado en la guía docente*" y 5 "*Las tutorías académicas con este/a profesor/a han resultado útiles*". Finalmente, con la función de información del *test* se determinó que el cuestionario de actividad docente proporciona mayor información sobre la satisfacción de los estudiantes que se encuentran en los niveles de rasgo medio - bajo del rasgo medido.

Un punto en común a tener en cuenta en la aplicación de los dos modelos corresponde al análisis de la fiabilidad. En la TCT utilizando el Coeficiente Alfa de Cronbach arrojó un indicador elevado (α=0.96) que está asociado directamente con la alta correlación entre los elementos de la escala. En algún momento de la investigación y dados los valores elevados del Alfa, se consideró que podría existir redundancia entre los ítems (Abad, Olea, Ponsoda y García, 2011). Sin embargo, al hacer lectura de estos se evidencia que cada uno de los ítems indaga por diferentes elementos que integran la actividad docente.

El coeficiente Omega mencionadas por Ventura-León y Caycho-Rodríguez, (2017), y por basarse en la estructura factorial que es más precisa cuando la escala es unidimensional, como es el caso del cuestionario de actividad docente, se obtuvo un indicador también elevado (ω=0.97), dato que se esperaba más elevado que en el alfa de Cronbach, de acuerdo con la subestimación que se da en la fiabilidad de los datos ordinales (Zumbo et al., 2007). En el caso de la TRI, esta medida de precisión se traslada a la Función de Información, que, a diferencia de la TCT, se estima para cada nivel de rasgo y no como un único indicador, en el caso del cuestionario de actividad docente se detectó que, si bien es elevada la precisión de la medida, esta se da principalmente para los niveles bajo - medio del rasgo. Adicionalmente, en los procedimientos basados en la TRI se estima para cada ítem un parámetro de discriminación y tantos parámetros b como el número de alternativas menos uno, además de la función de información del ítem y el *test* que permite realizar un análisis más detallado del ítem.

Es importante resaltar lo señalado por Penny (2003), en relación con la formación que reciben las personas que analizan y manejan la información resultante de los cuestionarios de actividad docente, así como la escasa capacitación que reciben sobre las posibilidades de análisis, por lo que es más frecuente encontrar análisis de tipo clásico, dado que en el análisis TRI se requiere comprender y desarrollar modelos matemáticos de mayor complejidad y con supuestos difíciles de cumplir.

Los indicadores obtenidos a partir del Modelo de Respuesta Graduada evidencian que con su uso es posible realizar un estudio más detallado de las propiedades de los ítems y que como ventaja adicional a la TCT permite seleccionar los ítems que se ajusten mejor a la función de información objetivo de la medición realizada. Ahora bien, funcionará mejor en la medida que el instrumento sea unidimensional, dado que no es necesario analizar las relaciones entre los factores. Otra ventaja del MRG es que informa sobre el comportamiento de cada ítem, por medio de las funciones

de respuesta para las categorías, con esta información y el análisis de su comportamiento en función del nivel de habilidad requerido para elegir una opción, es posible decidir sobre la pertinencia de tener en cuenta o no una categoría de respuesta. Estas características hacen que se recomiende el uso del MRG para el análisis de instrumentos de evaluación que usen una escala de respuesta graduada, tipo Likert, como la utilizada por los cuestionarios de evaluación docente. Entre las herramientas gráficas se resaltan las que se obtienen con el MRG, dado que en comparación con el modelo clásico que es más descriptivo; la función de información para cada ítem y para el *test* total facilitan la identificación de la precisión con la que se mide un atributo y en qué niveles de rasgo. Finalmente, como criterio práctico, el uso de modelos de TRI también permite proponer nuevos ítems que permitan realizar una medición precisa de los niveles de rasgo que se observan infraestimados. Por ejemplo, en el cuestionario analizado para esta universidad, los resultados señalan la necesidad de diseñar ítems que midan niveles altos de satisfacción con la evaluación docente.

Los análisis acá expuestos permiten evidenciar diferentes aspectos de los ítems que utilizados de forma complementaria permiten obtener mayor información de la calidad de los instrumentos utilizados en la evaluación docente, como señalan Barbero, Prieto, Suárez y San Luis (2001) y Kramp (2006), es posible establecer correspondencia entre las estimaciones realizadas con TCT e IRT y al analizar la información de cada técnica, se obtiene una mayor comprensión tanto de los sujetos como de los ítems. Estos dos modelos se pueden adecuar con facilidad para las diferentes instituciones universitarias que utilizan cuestionarios de evaluación docente con el fin de recolectar información sobre la satisfacción de los estudiantes y que se convierten en los principales indicadores de acreditación y como insumo principal en la toma de decisiones acerca de la labor del docente (Holland, 2019). Esto cobra relevancia, dado que, si el instrumento es utilizado para tomar decisiones sobre los docentes es fundamental conocer las características, ventajas y desventajas de dichos instrumentos. En ese sentido un instrumento de valoración confiable permitirá obtener información más precisa.

Un aspecto para reflexionar es que los modelos aplicados parten de un supuesto básico y es que existe una distribución de sujetos hacia un mismo objeto; sin embargo, la satisfacción de los estudiantes puede estar condicionada por los diferentes tipos de profesores que pueden existir en una universidad. Por lo que para futuras investigaciones se considera interesante analizar ¿cómo afecta al instrumento el hecho de que existan diferentes tipos de profesores? Si es tan fiable el instrumento en los niveles bajos de rasgo, quiere decir que la mayoría de las evaluaciones de los estudiantes son favorables y si son todas tan favorables, no parece que exista mucha variabilidad en los docentes; sin embargo, sería interesante analizar qué pasaría con los resultados y con la información que brinda el *test* si existiera una alta variabilidad en el profesorado que imparte los cursos. Otras líneas de investigación que pueden surgir a partir de estos resultados es comparar si las propiedades psicométricas del instrumento cambian en función del área en la que se desarrollan; por ejemplo, ciencias de la salud, ingenierías, ciencias sociales, entre otras. Así mismo, se considera importante indagar qué ocurre en la evaluación docente cuando se separan grupos de estudiantes, por ejemplo, comparar estudiantes de primer semestre y de último semestre, con el fin de identificar otros efectos que puedan influir en este tipo de evaluaciones o la necesidad de crear instrumentos acordes con diferentes grupos de estudiantes.

REFERENCIAS

Abad, F., Ponsoda, V. y Revuelta, J., (2006). *Modelos politómicos de respuesta al ítem*. Madrid: La Muralla.

Abad, F.; Olea, J.; Ponsoda, V. y García, C. (2011). *Medición en ciencias sociales y de la salud*. Madrid: Editorial Síntesis.

Abal, F. J. P., Lozzia, G. S., Aguerri, M. E., Galibert, M. S., & Attorresi, H. F. (2010). La escasa aplicación de la teoría de respuesta al ítem en ws de ejecución típica. *Revista Colombiana de Psicología, 19*(1), 111-122. Recuperado de https://www.redalyc.org/articulo.oa?id=80415077010

Abal, F. J. P., Auné, S. E., y Attorresi, H. F. (2014). Comparación del modelo de respuesta graduada y la Teoría Clásica de Tests en una escala de confianza para la matemática. *Summa Psicológica UST, 11*(2), pp. 101-113. Recuperado de https://summapsicologica.cl/index.php/summa/article/view/158

Asún, R., y Zúñiga, C. (2008). Ventajas de los modelos politómicos de teoría de respuesta al ítem en la medición de actitudes sociales: el análisis de un caso, P*sykhe* (Santiago)*, 17*(2), 103-115. Recuperado de https://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0718-22282008000200009

Attorresi, H., Abal, F., Galibert, M., Lozzia, G., y Aguerri, M. (2011). Aplicación del modelo de respuesta graduada a una escala de voluntad de trabajo. *Interdisciplinaria, 28*(2), pp. 231-244. Recuperado de https://www.academia.edu/es/30100086/Aplicaci%C3%B3n_del_Modelo_de_Respuesta_Graduada_a_una_Escala_de_Voluntad_de_Trabajo

Barbero, M. I., Prieto, P., Suárez, J. C. y San Luis, C. (2001). Relaciones empíricas entre los estadísticos de la teoría clásica de los tests y los de la teoría de respuesta a los ítems. *Psicothema, 13*(2), pp.324-329. Recuperado de https://dialnet.unirioja.es/servlet/articulo?codigo=2006890

Boring, A., Ottoboni, K., y Stark, P. (2016). Student evaluations of teaching (mostly) do not measure teaching effectiveness. *Science Open Research*, 1 - 11. Recuperado de https://www.math.upenn.edu/~pemantle/active-papers/Evals/stark2016.pdf

Cortina, J. M. (1993). What is coefficient alpha? an examination of theory and applications. *Journal of Applied Psychology, 78*, pp. 98-104. Recuperado de https://psycnet.apa.org/record/1993-19965-001

Denson, N., Loveday, T. y Dalton, H. (2010). Student evaluation of courses: What predicts satisfaction? *Higher Education Research and Development, 29*(4), pp.339–356. Recuperado de https://eric.ed.gov/?id=EJ888721

Elosua Oliden, P., y Zumbo, B. D. (2008). Coeficientes de fiabilidad para escalas de respuesta categórica ordenada. *Psicothema, 20*(4). Recuperado de https://dialnet.unirioja.es/ejemplar/201982

Espinosa, J., Fernández Sánchez, J. A., Tarí, J. J., Sabater Sempere, V., Valdés Conca, J., y García-Fernández, M. (2017). *Análisis de la calidad de la docencia en la universidad española*. Barcelona: Octaedro. Recuperado de https://rua.ua.es/dspace/handle/10045/71105?locale=es

Feistauer, D., y Richter, T. (2016). How reliable are students' evaluations of teaching quality? A variance components approach. Assessment & Evaluation in Higher Education, 42(8), pp. 1-17. Recuperado de https://www.researchgate.net/publication/311153533_How_reliable_are_students%27_evaluations_of_teaching_quality_A_variance_components_approach

Holland (2019) Making sense of module feedback: accounting for individual behaviours in student evaluations of teaching. *Assessment & Evaluation in Higher Education*, 44:6, 961-972, Recuperado de https://www.tandfonline.com/doi/abs/10.1080/02602938.2018.1556777

Kramp, U. (2006). *Efecto del número de opciones de respuesta sobre las propiedades psicométricas de los cuestionarios de personalidad* (Tesis doctoral). Barcelona: Universidad de Barcelona.

Martínez, R ., y Hernández, V . (2014) . *Psicometría*. Madrid, España: Difusora Larousse - Alianza Editorial

Montero, I., y León, O. G. (2002). Clasificación y descripción de las metodologías de investigación en Psicología. *International Journal of Clinical and Health Psychology, 2*(3), pp. 503-508. Recuperado de https://www.sciencedirect.com/search?qs=2002

Muñiz, J. (2010). Las teorías de los tests: teoría clásica y teoría de respuesta a los ítems. *Papeles del psicólogo, 31*, pp. 57-66. Recuperado de https://www.papelesdelpsicologo.es/pdf/1137.pdf

Muñiz, J. (2018). *Introducción a la psicometría*. Madrid: Pirámide.

Oermann, M. H., Conklin, J. L., Rushton, S., y Bush, M. A. (2018). Student evaluations of teaching (SET). *Guidelines for their use. Nursing forum, 53*(3), pp. 280-285. Recuperado de https://onlinelibrary.wiley.com/page/journal/17446198/homepage/forauthors.html

Penny, A.R. (2003). Changing the Agenda for Research into Students' Views about University Teaching: Four shortcomings of SRT research, *Teaching in Higher Education*, 8:3, 399-411, Recuperado de: https://www.tandfonline.com/doi/abs/10.1080/13562510309396

Revelle, W. (2017). *Psych: Procedures for Personality and Psychological Research*. Evanston, Illinois: Northwestern University. Recuperado de https://CRAN.R-project.org/package=psych Version = 1.7.8.

Samejima, F. (2010). The general graded response model. En Nering, M y Ostini, E. (Eds). Handbook of polytomous ítem responde models. New York: Taylor & Francis Group.

Spooren, P., Mortelmans, D., y Thijssen, P. (2012). 'Content'versus 'style': acquiescence in student evaluation of teaching? *British Educational Research Journal, 38*(1), pp. 3-21. Recuperado de https://bera-journals.onlinelibrary.wiley.com/toc/14693518/2012/38/1

Spooren, P., Brockx, B., y Mortelmans, D. (2013). On the validity of student evaluation of teaching: The state of the art. *Review of Educational Research, 83*(4), pp.598-642. Sage Publishing. Recuperado de https://journals.sagepub.com/home/rer

Stake, R. E., García, M. I. A., y Pérez, G. C. (2017). Evaluando la calidad de la Universidad–Particularmente su enseñanza. *REDU: Revista de Docencia Universitaria, 15*(2), pp. 125-142. Recuperado de https://polipapers.upv.es/index.php/REDU/issue/view/770

Tejedor, F. J. y Jornet, J. M. (2008). La evaluación del profesorado universitario en España. *Revista Electrónica de Investigación Educativa*, Especial. Recuperado de http://redie.uabc.mx/NumEsp1/contenidotejedorjornet.html

Turull, M., y Buxarrais, M. R. (2018). La evaluación de la docencia en las universidades públicas catalanas: análisis comparativo de los diferentes manuales de evaluación. *Revista de Educación y Derecho, 17*, pp. 1-30. Barcelona: Universidad de Barcelona. Recuperado de https://revistes.ub.edu/index.php/RED/article/view/21842

Ventura-León, J. L., y Caycho-Rodríguez, T. (2017). El coeficiente Omega: un método alternativo para la estimación de la confiabilidad. *Revista Latinoamericana de Ciencias Sociales, Niñez y Juventud, 15*(1), pp. 625-627. Manizales: Cinde. Recuperado de https://revistaumanizales.cinde.org.co/

Zumbo, B. D., Gadermann, A.M., y Zeisser, C. (2007). Ordinal versions of coefficients alpha and theta for likert rating scales. *Journal of Modern Applied Statistical Methods, 6*, pp.21-29. Recuperado de https://www.scirp.org/reference/ReferencesPapers.aspx?ReferenceID=2182035