Educación XX1
ISSN: 1139-613X · e-ISSN: 2174-5374

Medios e instrumentos para evaluar los resultados de aprendizaje en másteres universitarios. Análisis de la percepción del profesorado sobre su práctica evaluativa

Methods and instruments to assess learning outcomes in master's degrees. Analysis of teachers' perception of their evaluative practice

María Soledad Ibarra-Sáiz [1*]
Gregorio Rodríguez-Gómez [1]
José Francisco Lukas-Mujika [2]
Alaitz Santos-Berrondo [2]

[1] Universidad de Cádiz, Spain
[2] Universidad del País Vasco/Euskal Herriko Unibertsitatea (UPV/EHU), Spain
* Autor de correspondencia. E-mail: Marisol.ibarra@uca.es

Cómo referenciar este artículo/How to reference this article:

Ibarra-Sáiz, M.S., Rodríguez-Gómez, G., Lukas-Mujika, J.F., & Santos-Berrondo, A. (2023). Medios e instrumentos para evaluar los resultados de aprendizaje en másteres universitarios. Análisis de la percepción del profesorado sobre su práctica evaluativa. [Methods and instruments to assess learning outcomes in master's degrees. Analysis of teachers' perception of their evaluative practice]. *Educación XX1, 26*(1), 21-45. https://doi.org/10.5944/educxx1.33443

Fecha de recepción: 14/03/2022
Fecha de aceptación: 13/09/2022
Publicado online: 02/01/2023

RESUMEN

Estudios previos sobre los medios e instrumentos de evaluación utilizados en la educación superior han puesto de manifiesto el uso mayoritario del examen final como principal fuente de valoración. Los avances en el conocimiento de los procesos de evaluación han evidenciado la necesidad de disponer de una mayor amplitud y diversidad de medios e instrumentos que permitan recabar una información rigurosa y válida sobre la que sustentar los juicios sobre el grado de aprendizaje del estudiantado. Este estudio se ha realizado en el contexto del Proyecto FLOASS (http://floass.uca.es) con la finalidad de explorar la percepción que sobre su práctica evaluativa tiene el profesorado. Se ha utilizado una metodología mixta, mediante un diseño secuencial exploratorio, que ha permitido recabar la percepción de 416 profesores, de seis universidades de diferentes comunidades autónomas, que cumplimentaron el cuestionario RAPEVA-*Autoinforme del profesorado sobre su práctica en la evaluación de los resultados de aprendizaje*. Entre los medios más utilizados destaca la participación, las pruebas de resolución de problemas, pruebas de desempeño, objetos digitales o presentaciones multimedia y los proyectos y las rúbricas o el argumentario evaluativo entre los instrumentos de evaluación. Se han encontrado las mayores diferencias en función de la universidad, el ámbito de conocimiento o el grado de seguridad y satisfacción con el sistema de evaluación. En el caso del género o la experiencia las diferencias son menores o inexistentes. Se aportan futuras líneas de investigación que posibiliten una mayor comprensión de la práctica evaluativa en la educación superior.

Palabras clave: educación superior, evaluación, evaluación del estudiante, evaluación formativa, evaluación sumativa

ABSTRACT

Previous studies on the assessment methods and instruments used in higher education have revealed that the final exam has been widely used as the main source of assessment. Advances in knowledge of assessment processes have shown the need to have a greater breadth and diversity of methods and instruments that allow the collection of thorough and valid information on which to base judgments about the level of learning in students. Within the framework of the FLOASS Project, this study has been carried out in order to explore the perception that teachers have of their assessment practice. A mixed methodology has been used, through an exploratory sequential design, which has allowed to gather the perception of 416 professors from six universities belonging to different autonomous communities, who completed the RAPEVA questionnaire – *Self-report of the teaching staff on their practice in the learning outcomes assessment.* Among the most widely used methods, participation, problem solving tests, performance tests, digital objects or multimedia presentations and projects and rubrics or evaluative arguments are highlighted among the assessment instruments. The greatest differences were found depending on the university, the field of knowledge or the degree of security and satisfaction with the assessment system. In the case of gender or experience, differences are small or non-existent. Future lines of research that enable a better understanding of assessment practice in higher education are provided.

Keywords: higher education, assessment, student assessment, performance assessment, summative assessment

INTRODUCCIÓN

Los medios e instrumentos utilizados para evaluar el aprendizaje de los estudiantes en la universidad son una parte importante del proceso de enseñanza-aprendizaje dado que la calidad de los mismos va a garantizar o no la consecución de unos resultados de aprendizaje alineados con procesos cognitivos superiores. Cuando Sadler (2016) advertía que no debemos confundir una evidencia de baja calidad del rendimiento del estudiante con la evidencia de un bajo rendimiento, destacaba la enorme importancia que tienen los medios e instrumentos de evaluación, pues solo en la medida en que estos sean válidos, pertinentes y adecuados podremos realizar inferencias justas y fundamentadas sobre el nivel de rendimiento de los estudiantes.

El trabajo de Ibarra-Sáiz y Rodríguez-Gómez (2010) evidenció la preponderancia del clásico examen como medio esencial para la evaluación en la educación superior. Posteriormente, el estudio de Rodríguez-Gómez et al. (2013) mostró cierta evolución en el uso de otros medios e instrumentos de evaluación. Más recientemente, el trabajo de Panadero et al. (2019) da señales de que se ha producido un lento y progresivo cambio en la diversidad de los medios e instrumentos de evaluación utilizados en la educación superior, aunque sigue constatándose la prevalencia de un examen final como la fuente de información esencial para determinar las calificaciones finales.

Estos estudios previos se sustentaron básicamente sobre la base del análisis documental de las guías docentes o programas de las asignaturas, y era preciso profundizar en este aspecto a partir de otras fuentes de información que aportaran una perspectiva actual y diferente. El estudio que se presenta a continuación se centra, específicamente, en la percepción del profesorado, y se ha realizado en el contexto más amplio que constituye el Proyecto FLOASS (Ibarra-Sáiz & Rodríguez-Gómez, 2019). En este proyecto se pretende ofrecer un marco de acción, sustentado en el uso de tecnologías que mejoran la evaluación (*Technology Enhance Assessment*-TEA) y analíticas de aprendizaje (*Learning Analytics*-LA), que oriente el diseño, implementación, seguimiento y evaluación de resultados de aprendizaje (*Learning Outcomes*-LO) que exijan altas capacidades de los estudiantes.

Este estudio parte del análisis de la perspectiva del propio profesorado universitario, expresada esta a través de un autoinforme individual y circunscrito al contexto de los másteres en el área de las ciencias sociales. Con ello se pretendía explorar la práctica evaluativa del profesorado universitario, a partir de sus propias percepciones, centrando la atención de forma concreta en uno de

los múltiples elementos constitutivos de esta práctica evaluativa, como son los medios e instrumentos de evaluación que se utilizan para evaluar los resultados de aprendizaje.

Analizar la percepción del profesorado es esencial para comprender y mejorar la práctica evaluativa. Una comprensión más profunda de estas prácticas permitirá centrar la atención en la mejora del aprendizaje del estudiante y, en consecuencia, en aquellos aspectos que faciliten el cambio y la innovación en la enseñanza. Los estudios previos de Ibarra-Sáiz y Rodríguez-Gómez (2014) y Panadero et al. (2019) destacan la importancia de variables contextuales como la universidad de procedencia, pero es preciso analizar otras variables contextuales como el ámbito de conocimiento, o personales como el género, la experiencia o la seguridad y satisfacción evaluadora autopercibida, que no han sido consideradas previamente.

Medios e instrumentos para la evaluación en educación superior

En el campo de la evaluación educativa en educación superior se han producido avances y cambios de los que Boud (2020) destaca, en primer lugar, que las políticas de evaluación en la educación superior se basen menos en normas y más en principios, otorgándose así una mayor flexibilidad en los procesos de evaluación. En segundo lugar, el incremento de la evaluación auténtica, que implica el uso de tareas y procesos de evaluación representativos de los tipos de tareas y procesos que se encuentran en la práctica profesional. A pesar de estos cambios, como señala este autor, "un extraño se sorprendería al ver cuánta práctica que no puede defenderse sobre la base de ningún conocimiento evaluativo sigue existiendo" (Boud, 2020, p. 8).

Un ejemplo de este cierto desconocimiento nos lo encontramos al abordar el análisis de los medios e instrumentos de evaluación, ya que nos enfrentamos con una dificultad arraigada en la confusión conceptual y terminológica que circunda a estos dos conceptos evaluativos, que llegan a utilizarse como sinónimos o equivalentes. Así, Mateo Andrés y Martínez Olmo (2008) presentan una serie de procedimientos evaluativos alternativos que llegan a equiparar a ejecuciones o actividades. En este estudio se partía de una clara diferenciación conceptual, planteada en su momento por Rodríguez-Gómez e Ibarra-Sáiz (2011), quienes consideran que "los medios de evaluación son las pruebas o evidencias que sirven para recabar información sobre el objeto a evaluar" (p. 71), es decir, los productos o actuaciones que realiza el estudiantado; y los instrumentos de evaluación son las herramientas que utiliza el evaluador para realizar de una forma sistemática sus valoraciones sobre los múltiples y diferentes aspectos o características susceptibles de valoración en un producto o actuación del estudiantado. Así, un portafolio,

una exposición oral, un ensayo o el informe de una práctica de laboratorio son claros ejemplos de productos o actuaciones del estudiantado que se constituyen o transforman en medios de evaluación para quienes tengan que valorar su calidad, ya sea el profesorado, los compañeros a través de la evaluación entre iguales, el propio estudiante a través de la autoevaluación, o conjuntamente profesorado y estudiantado (coevaluación).

El análisis y la valoración sistemática de estos medios de evaluación (productos y actuaciones del estudiante) exigen instrumentos que faciliten la emisión de los juicios de valor del evaluador. Así, por ejemplo, contamos con instrumentos como las listas de control, que permitirán la valoración de la presencia o ausencia de determinadas características; las escalas de estimación, con las que se podrán valorar el mayor o menor grado en el que se presentan esas características a valorar; las rúbricas, que permitirán valorar de una forma más descriptiva y exhaustiva diferentes niveles de logro, o el argumentario evaluativo en el que se realizan valoraciones cualitativas.

Los cambios de rumbo en la evaluación educativa destacados por Boud (2020) suponen una modificación en los medios de evaluación, que han pasado del uso de las clásicas pruebas, test o exámenes finales centrados en la reproducción del conocimiento, a un conjunto de medios de evaluación (portafolio, simulaciones, resolución de casos, …), a través de los cuales se pretende integrar y dar coherencia al aprendizaje que se pretende desarrollar (Ibarra-Sáiz & Rodríguez-Gómez, 2020. Como ya señaló Dochy (2009), estos medios de evaluación centran el foco en las ejecuciones del estudiante, en lo que produce y hace, en lo que es capaz de realizar y producir, utilizando el pensamiento crítico y la creatividad en el momento de resolver problemas, y que estos sean actuales.

El alineamiento constructivo del currículo (Biggs, 2015; Biggs & Tang, 2011) trata de asegurar la utilidad del proceso de evaluación, que lo que se enseñe sea relevante y que esté orientado a los resultados de aprendizaje. Desde esta consideración, y tomando como base el concepto de competencia del European Centre for Development of Vocational Training (CEDEFOP, 2014), expresado como la capacidad de una persona para poner en práctica adecuadamente los resultados de aprendizaje en un contexto concreto (educación, trabajo o desarrollo personal o profesional), la evaluación del desarrollo competencial del estudiantado exige la disponibilidad de pruebas y evidencias de en qué grado se han alcanzado los resultados de aprendizaje deseados (Brown & Pickford, 2013) lo que exige disponer de medios e instrumentos de evaluación que evidencien de manera clara la complejidad cognitiva, que sean equitativos, que favorezcan la transparencia de los procedimientos de evaluación y que sean útiles para el aprendizaje del estudiantado (Dochy, 2009).

OBJETIVOS Y CUESTIONES DE INVESTIGACIÓN

El objetivo de este trabajo se centró en explorar la práctica evaluativa del profesorado universitario a partir de sus propias percepciones, centrando la atención de forma específica en los medios e instrumentos de evaluación utilizados para evaluar los resultados de aprendizaje. Las cuestiones de investigación que lo han orientado han sido:

- ¿Qué medios e instrumentos de evaluación suele utilizar el profesorado en su práctica evaluativa?
- ¿Existe alguna diferencia en el uso de medios e instrumentos de evaluación que pueda asociarse a características como la universidad de pertenencia, el ámbito de conocimiento, el género, los años de experiencia o la seguridad y satisfacción evaluadora autopercibida?
- En comparación con estudios previos, ¿se denota alguna evolución en los medios e instrumentos que se utilizan para la evaluación en el contexto universitario?

MÉTODO

Diseño de investigación

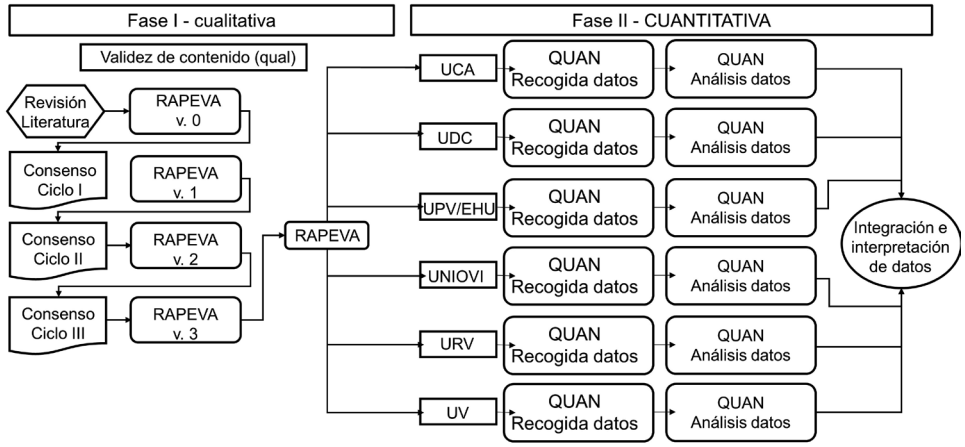
Para realizar este estudio se ha optado por una metodología mixta que se concreta en el diseño que Creswell (2015) denomina secuencial exploratorio, en el que el énfasis se sitúa en la fase cuantitativa (qual->QUAN) (Figura 1). En la primera fase de la investigación se procedió al diseño y validación de contenido del cuestionario RAPEVA-*Autoinforme del profesorado sobre su práctica en la evaluación de los resultados de aprendizaje*. En la segunda fase se procedió, mediante un proceso de encuestación, a recabar las percepciones de profesores universitarios de diferentes regiones autónomas del estado español. La recogida de datos se realizó durante el segundo semestre del curso académico 2020/2021.

El autoinforme RAPEVA

La construcción del autoinforme RAPEVA se inició con una revisión de la literatura y, posteriormente, se procedió a un proceso de validación mediante jueces (Figura 1). En la misma participaron 22 jueces a los que se les pidió que valoraran cada uno de los ítems en función de la congruencia (la afirmación mide realmente la dimensión en la que está clasificada), la claridad (está bien redactada y es comprensible) y la relevancia (si es importante para medir la dimensión en la

Figura 1

Diseño secuencial exploratorio



que se incluye). De los diferentes métodos de validación de contenido (Johnson & Morgan, 2016), se optó por el método de consenso grupal, evitando así los sistemas de votación. Para ello, se realizaron diversas reuniones con los expertos hasta llegar en cada caso al citado consenso, aspecto que no resultó difícil dada la alta valoración que tuvieron la mayoría de los ítems en los tres indicadores citados. La definición y concreción de los diferentes indicadores se revisaron al finalizar cada uno de los ciclos.

El autoinforme RAPEVA se basa en la modelización de constructos como compuestos, es decir, como combinaciones lineales de las variables observadas (Henseler, 2021), ya que, en la valoración que el profesorado realice de cada uno de los ítems, juegan un papel esencial los aspectos cognitivos y actitudinales, constituyéndose así en un índice formativo (Hair et al, 2022). Por ello, se ha optado por realizar un análisis generalizado de componentes estructurales (Hwang & Takane, 2015), obteniendo medidas de ajuste (GFI=.89 y SRMR=.08) que se consideran aceptables.

Inicialmente se solicita información sobre aspectos contextuales como la universidad de procedencia, el ámbito de conocimiento, años de experiencia o el género. En la segunda se presentan 49 ítems en formato de escala tipo Likert (0-5) estructurados en diez dimensiones (Tabla 1). La cumplimentación del autoinforme requería unos 20 minutos.

Este estudio se centra, únicamente, en las cuatro dimensiones relacionadas con los medios e instrumentos de evaluación (MOB, MEN, MDA e INE), dejando las referentes a las tareas de evaluación para su presentación y divulgación en un estudio diferente.

Tabla 1
Estructura del autoinforme RAPEVA

	Dimensiones	# Items	Items
TRA	Transparencia	5	I01, I03 al I05, I35
CAE	Competencias a evaluar	6	I06 al I11
MOB[a]	Medios de observación	6	I12 al I17
MEN[a]	Medios de encuesta	4	I18 al I21
MDA[a]	Documentos y artefactos	9	I22 al I30
INE[a]	Instrumentos de evaluación	4	I31 al I34
PRO	Profundidad de las tareas	4	I36 al I38
RET	Retroalimentación	3	I39 al I41
PAR	Participación	4	I42 al I45
FOR	Formación en evaluación	2	I46, I47
SSE	Satisfacción con evaluación	2	I48, I49

[a] Dimensiones objeto de este estudio.

Participantes

En la fase cualitativa (validación de contenido) intervinieron los 22 miembros de los equipos de investigación de las seis universidades participantes en el proyecto. Una vez validado el autoinforme, en la fase cuantitativa se invitó a todo el profesorado que impartía clases en másteres de Ciencias Sociales de las seis universidades, mediante un correo electrónico en el que se indicaba el enlace a través del cual cumplimentar el autoinforme. Esta invitación al profesorado se realizó a través de correo electrónico por parte de la coordinación de cada uno de los másteres. Cada docente que recibía el correo, lo hacía por participar en un máster específico en el que impartía una asignatura. Sus respuestas debían ser siempre pensando en esa asignatura y no en otras que pudiera impartir en otras titulaciones. Iniciaron la cumplimentación un total de 626 profesores, finalmente se obtuvieron 416 autoinformes completos del profesorado que impartía docencia en 63 másteres (Tabla 2), de los que el 47.6% eran del ámbito de la educación (EDU), el 44.4% se correspondían con economía y empresa (ECO) y el 7.9% con comunicación (COM).

Tabla 2
Distribución de másteres por universidades y ámbito de conocimiento

Ámbito	UCA	URV	UNIOVI	UV	UDC	UPV/EHU	Total
COM	1	1			1	2	5
ECO	7	1	2	6	5	7	28
EDU	4	4	4	7	6	5	30
Total	12	6	6	13	12	14	63

Tabla 3
Características demográficas

	Mujer		Hombre		Otros		Total	
Universidad	n	%	n	%	n	%	n	%
UCA	30	49.2	31	50.8	0	0	61	14.7
URV	18	60	12	40	0	0	30	7.2
UNIOVI	35	63.6	20	36.4	0	0	55	13.2
UV	36	51.4	33	47.1	1	1.4	70	16.8
UDC	45	54.2	37	44.6	1	1.2	83	20
UPV/EHU	61	52.1	56	47.9	0	0	117	28.1
Ámbito								
COM	10	55.6	8	44.4	0	0	18	4.3
ECO	83	49.7	82	49.1	2	1.2	167	40.1
EDU	132	57.1	99	42.9	0	0	231	55.5
Experiencia								
<10 años	59	64.8	32	35.2	0	0	91	21.9
11-20 años	78	64.5	41	33.9	2	1.7	121	29.1
>20 años	88	43.1	116	56.9	0	0	204	49
Satisfacción								
Nivel 1	75	68.2	34	30.9	1	0.9	110	26.4
Nivel 2	104	59.1	72	40.9	0	0	176	42.3
Nivel 3	8	26.7	22	73.3	0	0	30	7.2
Nivel 4	38	38	61	61	1	1	100	24
Total	225	54.1	189	45.4	2	0.5	416	100

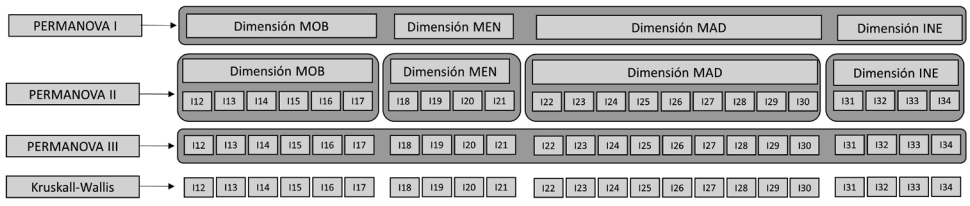
En la Tabla 3 se presenta la distribución de los 416 profesores en función de la universidad de procedencia, el género, ámbito de conocimiento en el que impartían clases (comunicación (COM), economía y empresa (ECO) y educación (EDU), años de experiencia, y grado de satisfacción y seguridad con la evaluación. Para construir los niveles de satisfacción y seguridad se tuvieron en cuenta las percepciones de los propios docentes en sus respuestas a los ítems 48 y 49 del cuestionario y, sobre la base de las puntuaciones obtenidas, se establecieron cuatro niveles en función de los cuartiles.

Análisis de datos

Para dar respuesta al primer interrogante, se realizó una exploración con diagramas de cajas y bigotes y estadísticos descriptivos de tendencia central y dispersión. En segundo lugar, para analizar las diferencias entre grupos (segundo interrogante), se han realizado pruebas no paramétricas, ya que se trataban de mediciones ordinales que no se ajustaban a la normalidad (Prueba de K-S, $p<.001$). En la Figura 2 se presenta el proceso de arriba-abajo que se ha seguido en el análisis comparativo, utilizando para ello el *Permutational Multivariate Analysis of Variance* (PERMANOVA) en tres momentos diferentes. Esta técnica de análisis multivariante (sobre medidas de distancia) con varios factores aplica el análisis de permutaciones sobre las matrices de distancias (Anderson, 2017) para establecer la comparación multivariante. En primer lugar, se contrastaron las posibles diferencias entre grupos comparando las cuatro dimensiones globales (PERMANOVA I); para profundizar en las posibles diferencias encontradas se compararon individualmente los ítems que constituyen cada una de las dimensiones (PERMANOVA II); y por último, para comprender las relaciones entre las variables se realizó un último análisis multivariante con todos los ítems a la vez (PERMANOVA III). Finalmente se procedió al análisis individual de cada ítem mediante la prueba H de Kruskal-Wallis. Para la ejecución de estos análisis se utilizaron JASP (JASP Team, 2022) y R (R Core Team, 2021).

Figura 2

Proceso del análisis comparativo



RESULTADOS

Medios e instrumentos de evaluación utilizados

En respuesta al primer interrogante de este estudio, en la Figura 3 se observa la tendencia central (mediana), dispersión (rango intercuartílico) y asimetría de las valoraciones realizadas en cada una de las cuatro dimensiones. En la Tabla 4 se presentan las medidas de tendencia central y dispersión por dimensiones e ítems. El mayor grado de acuerdo o frecuencia se produce en la dimensión de los medios de observación (M=2.84), y en un nivel inferior se encuentran las dimensiones de los medios de encuestación (M=2.22) y documentos y artefactos (M=2.08). En los instrumentos de evaluación se alcanza una M=2.21.

Figura 3

Diagrama de caja y bigotes correspondiente a las dimensiones del autoinforme RAPEVA

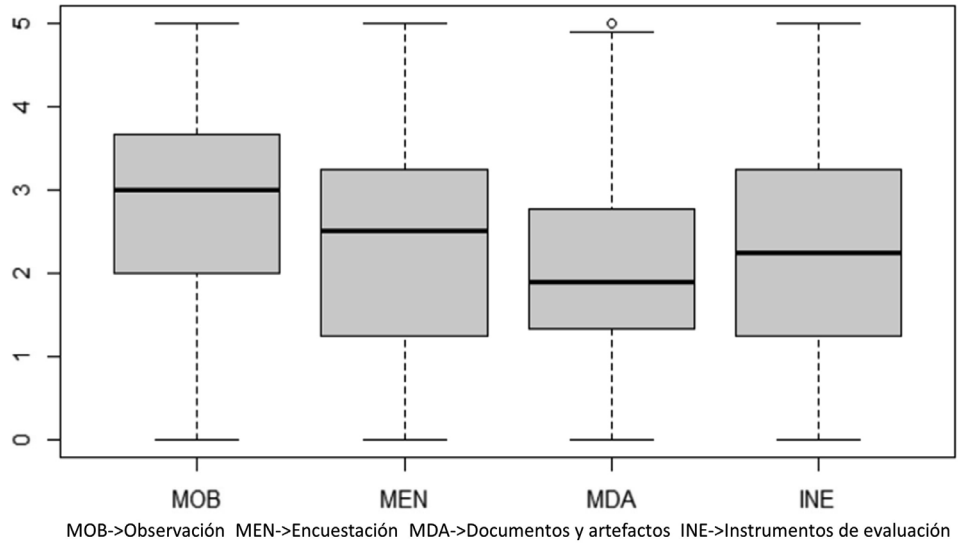


Tabla 4
Medidas de tendencia central y dispersión en cada ítem

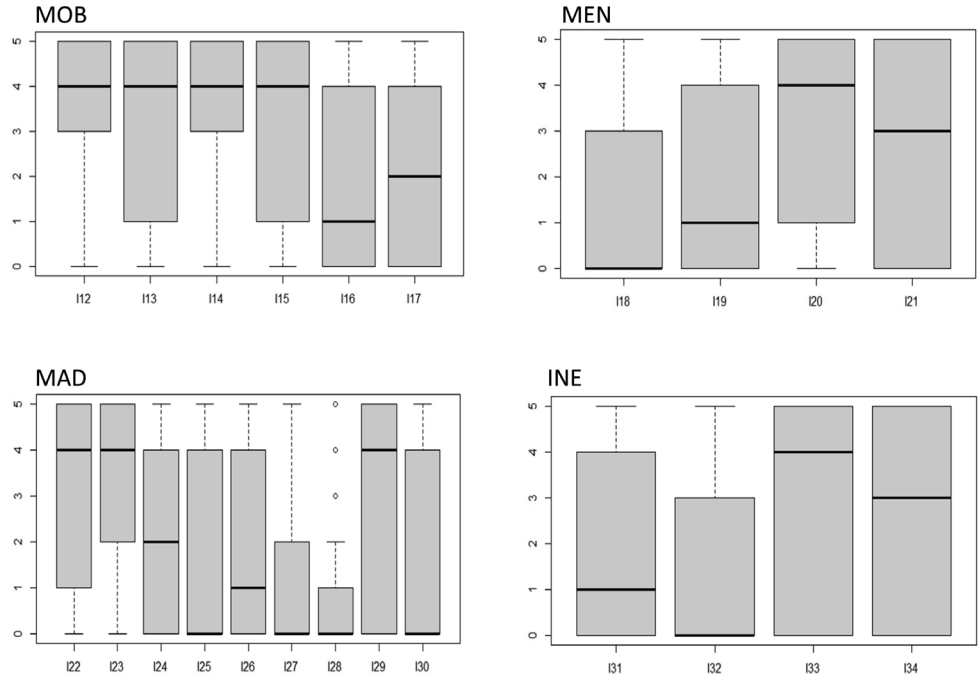
	Mdn	M	SD
Observación (MOB)	3	2.84	1.16
I12. Participación activa de los estudiantes en actividades	4	3.72	1.50
I13. Prácticas del alumnado	4	2.95	1.95
I14. Pruebas de resolución de problemas	4	3.43	1.71
I15. Pruebas de desempeño	4	2.92	1.89
I16. Talleres	1	1.88	1.92
I17. Simulaciones	2	2.17	2.01
Encuestación (MEN)	2.5	2.22	1.42
I18. Entrevistas orales	0	1.35	1.80
I19. Grupos de discusión o grupos focales	1	1.93	2.00
I20. Pruebas de resolución de problemas	4	2.97	2.00
I21. Pruebas de desempeño	3	2.63	2.07
Análisis de documentos y artefactos (MDA)	1.89	2.08	1.04
I22. Objetos digitales o presentaciones multimedia	4	3.16	1.96
I23. Pruebas de resolución de problemas	4	3.15	1.87
I24. Pruebas objetivas	2	2.12	2.13
I25. Pruebas escritas con preguntas de respuesta breve	0	1.64	2.00
I26. Pruebas escritas de desarrollo	1	2.05	2.16
I27. Pruebas escritas de mapas conceptuales	0	.96	1.60
I28. Diarios	0	.96	1.67
I29. Proyectos	4	3.16	2.12
I30. Portafolio	0	1.55	2.05
Instrumentos de evaluación (INE)	2.25	2.21	1.28
I31. Listas de control o verificación	1	2.01	2.09
I32. Escalas de estimación	0	1.30	1.85
I33. Rúbricas	4	2.87	2.15
I34. Argumentario valorativo	3	2.64	2.10

Medios utilizados en la observación (MOB)

La observación es utilizada para la evaluación de los productos o actuaciones de los estudiantes en algunas de las actividades que realizan (Tabla 4 y Figura 4). Así, por ejemplo, el 68% del profesorado señala que mediante la observación comprueba la participación activa de los estudiantes frente al 11.6% que nunca o casi nunca la utiliza. También la observación es utilizada para analizar las pruebas de resolución de problemas (61.5%), prácticas del alumnado (51.5%), y pruebas de desempeño (52.2%). Sin embargo, su uso es mucho menor en las simulaciones (33.4%) y talleres (27.9%).

Figura 4

Diagrama de caja y bigotes de los ítems en llas dimensiones MOB, MEN, MAD e INE



Medios utilizados en la encuestación (MEN)

En este caso las respuestas han sido desiguales. Se ha comprobado el uso medio de las pruebas de resolución de problemas, dado que el 54.8% del profesorado afirma utilizarlas frente a casi el 30% que nunca o casi nunca lo hace. Igualmente, las pruebas de desempeño se utilizan por el 47.9%. Sin embargo,

son menos habituales el uso de los grupos de discusión o grupos focales y las entrevistas orales, puesto que son utilizadas habitualmente solo por el 31% y el 19.2% respectivamente.

Medios utilizados para el análisis de documentos y artefactos (MAD)

Se constata un uso desigual en esta categoría (Tabla 4 y Figura 4). El profesorado manifiesta un uso medio de los proyectos (60.1%), objetos digitales o presentaciones multimedia (57.9%), y pruebas de resolución de problemas (56.5%). Tienen un uso más limitado las pruebas objetivas (36.3%) y las pruebas escritas de desarrollo (35.9%). Aún más bajo es el uso de pruebas escritas con preguntas de respuesta breve (26.5%), y el portafolio (26.2%). Por último, hay un uso casi residual de los diarios (13.5%) y de las pruebas escritas de mapas conceptuales (12.3%).

Instrumentos de evaluación (INE)

Se comprueba en la Tabla 4 y en la Figura 4 que se manifiesta un uso medio de las rúbricas (53.3%) y del argumentario valorativo (49%). El uso de las listas de control (34.8%) y escalas de estimación (20.4%) es bajo.

Diferencias en el uso de medios e instrumentos de evaluación

Respecto al segundo interrogante planteado en este estudio, en la Tabla 5 se presentan los resultados obtenidos de los análisis PERMANOVA. Siendo variables dependientes las cuatro dimensiones (PERMANOVA I), se encontraron diferencias significativas en función de la universidad ($F_{5, 415}$=4.59, p=.001), el ámbito de conocimiento ($F_{2, 415}$=7.85, p=.001), y la seguridad y satisfacción con la evaluación ($F_{3, 415}$=3.02, p=.010), no encontrándose diferencias ni por género ($F_{2, 415}$=1.86, p=.091) ni por experiencia docente ($F_{2, 415}$=1.37, p=.233).

Al considerar como variables dependientes los ítems que constituían cada una de las cuatro dimensiones (PERMANOVA II) se encontraron diferencias significativas en función de la universidad y el ámbito de conocimiento. En función del género y de la experiencia docente, se encontraron diferencias significativas al considerar los ítems que conformaban la dimensión instrumentos de evaluación.

En el PERMANOVA III se consideraron como variables independientes los 23 ítems y, en este caso, las diferencias se encontraron en función de la universidad (p=.001), el ámbito de conocimiento (p=.001), la experiencia docente (p=.027) y la seguridad y satisfacción (p=.002).

Por último, para analizar cada uno de los ítems de forma independiente, se realizó la prueba H de Kruskall-Wallis (Tabla 6).

Diferencias en función de la universidad

En 15 de los 23 ítems se han encontrado diferencias significativas ($p \leq .05$). En la observación se evidencian diferencias en todos los medios de evaluación, excepto en la observación de las prácticas. En este sentido, aunque es difícil vislumbrar pautas de actuación de las distintas universidades, tras el análisis comparativo, es la UPV/EHU en general, en la que se manifiesta una menor utilización en casi todos los casos. La URV (M=4.37) y la UCA (M=4.26) destacan en la observación de la participación activa de los estudiantes. También la UCA (M=3.72), junto con la UDC (M=3.64) y UV (M=3.63) sobresale en la observación de la resolución de problemas. En las pruebas de desempeño, las puntuaciones de UDC (M=3.40), UV (M=3.21), UNIOVI (M=3.02) y UCA (M=2.98) son más elevadas que las de UPV/EHU (M=2.55) y URV (M=2.07). En la observación de los talleres, aunque las puntuaciones son bajas en general, sobresalen las de la UDC (M=2.64) frente al resto de las universidades. Por último, en la observación de las simulaciones, también se dan diferencias significativas entre la UV (M=3.14) y la UPV/EHU (M=1.56).

Estas diferencias entre universidades también se constatan en la utilización de las entrevistas orales y las pruebas de desempeño. En ambos casos, el profesorado de la UPV/EHU (M=0.97) y la URV (M=0.63) manifiestan una menor utilización que el profesorado de las restantes universidades.

Igualmente se constata en el análisis comparativo que el profesorado de la UPV/EHU (M=2.65) manifiesta un menor uso de objetos digitales o presentaciones multimedia que el de las demás universidades. Las pruebas objetivas son más utilizadas por el profesorado de la UDC (M=2.76) y la UCA (M=2.48), mientras que apenas se usan en la UPV/EHU (M=1.79) y en la URV (M=0.67). Por el contrario, las pruebas escritas de desarrollo son más utilizadas en la UV (M=2.99) y apenas se utilizan en la URV (M=0.67). Aunque el uso del portafolios está poco extendido en general, el profesorado de la UPV/EHU (M=1.11) y la URV (M=0.70) son los que expresan una menor utilización.

Respecto a los instrumentos de evaluación, se manifiesta un mayor grado de utilización por parte del profesorado de la UCA (M=1.90) y la UNIOVI (M=1.75) frente al de la UPV/EHU (M=0.91) y la URV (M=0.60). En el uso de las rúbricas sobresale el profesorado de la URV (M=4.07) frente al de las demás universidades. Y en el uso del argumentario evaluativo, es el profesorado de la UV (M=3.19) y URV (M=2.97) quienes expresan un mayor grado de utilización.

Tabla 5

Resultados PERMANOVA en función de universidad (UNI), ámbito de conocimiento (AMB), género (GEN), experiencia (EXP) y satisfacción (SSE)

		UNI		AMB		GEN		EXP		SSE	
Dimensiones		F	Sig.	F	Sig.	F	Sig.	F	Sig.	F	Sig.
Dimensiones globales	PERMANOVA I	4.5876	.001*	7.8502	.001*	1.8627	.091	1.3656	.233	3.0204	.010*
Ítems Dimensión MOB	PERMANOVA II	3.9662	.001*	7.6496	.001*	1.1733	.304	1.8116	.077	2.1547	.016*
Ítems Dimensión MEN		2.5391	.003*	3.0045	.011*	0.7894	.598	0.9191	.461	2.3929	.019*
Ítems Dimensión MDA		2.8245	.001*	7.2441	.001*	1.4389	.114	1.6208	.067	2.1283	.006*
Ítems Dimensión INE		2.6594	.001*	7.3225	.001*	2.1433	.035*	2.7198	.009*	1.41	.181
Ítems RAPEVA	PERMANOVA III	3.0041	.001*	6.5865	.001*	1.3968	.105	1.7511	.027	2.0421	.002*

*p≤.05

Tabla 6

Resultados de la prueba H de Kruskall Wallis en función de la universidad (UNI), ámbito de conocimiento (AM), género (GEN), experiencia (EXP) y satisfacción evaluadora (SSE)

	UNI		AMB		GEN		EXP		SSE	
	H	Sig.	H	Sig.	H	Sig.	H	Sig.	H	Sig.
Observación (MOB)										
I12. Participación activa de los estudiantes en actividades	22.443	.000*	3.832	.147	7.820	.020*	.819	.664	21.343	.000*
I13. Prácticas del alumnado	6.156	.291	19.571	.000*	1.136	.567	.1183	.912	8.296	.040*
I14. Pruebas de resolución de problemas	11.552	.041*	3.714	.156	.048	.976	.540	.763	10.095	.018*
I15. Pruebas de desempeño	17.904	.003*	.105	.949	.002	.999	1.699	.428	8.554	.036*
I16. Talleres	24.289	.000*	28.657	.000*	2.633	.268	2.789	.248	2.946	.400
I17. Simulaciones	34.951	.000*	25.938	.000*	2.744	.254	11.404	.003*	2.154	.541
Encuestación (MEN)										
I18. Entrevistas orales	15.000	.010*	17.486	.000*	1.446	.485	.675	.714	7.832	.050*
I19. Grupos de discusión o grupos focales	9.101	.105	10.257	.006*	3.187	.203	.605	.739	7.973	.047*
I20. Pruebas de resolución de problemas	8.881	.114	1.859	.395	2.227	.328	3.167	.205	9.682	.021*
I21. Pruebas de desempeño	13.632	.018*	1.471	.479	.053	.974	2.795	.247	6.858	.077
Análisis de documentos y artefactos (MDA)										
I22. Objetos digitales o presentaciones multimedia	14.504	.013*	10.909	.004*	9.160	.010*	7.520	.023*	10.417	.015*
I23. Pruebas de resolución de problemas	11.717	.039*	.052	.974	.030	.985	1.957	.376	6.873	.076
I24. Pruebas objetivas	26.740	.000*	10.512	.005*	2.224	.329	.213	.899	19.760	.000*
I25. Pruebas escritas con preguntas de respuesta breve	10.196	.070	7.143	.028*	1.041	.594	1.398	.497	4.611	.203
I26. Pruebas escritas de desarrollo	27.681	.000*	.341	.843	.889	.641	.118	.943	5.991	.112
I27. Pruebas escritas de mapas conceptuales	4.750	.447	13.426	.001*	1.967	.374	2.847	.241	1.935	.586
I28. Diarios	5.262	.385	23.734	.000*	2.938	.230	1.075	.584	1.907	.592
I29. Proyectos	6.163	.291	13.144	.001*	5.234	.073	7.261	.026*	4.364	.225
I30. Portafolio	19.668	.001*	43.264	.000*	1.955	.376	.357	.837	5.684	.128
Instrumentos de evaluación (INE)										
I31. Listas de control o verificación	10.213	.069	3.948	.139	2.961	.228	4.202	.122	2.111	.550
I32. Escalas de estimación	20.260	.001*	15.190	.001*	2.669	.263	.027	.986	4.776	.189
I33. Rúbricas	17.541	.004*	20.085	.000*	8.965	.011*	14.225	.001*	3.939	.268
I34. Argumentario valorativo	11.241	.047*	10.736	.005*	2.180	.336	.061	.970	8.055	.045*

* p≤.05

Diferencias en función del ámbito de conocimiento

También son 15 los ítems en los que se han detectado diferencias significativas. En la dimensión de observación, estas se evidencian en tres de los medios de evaluación. Tanto en el uso de las prácticas del alumnado (M=3.61) como en la de los talleres (M=3.33), es el profesorado de comunicación el que expresa una mayor utilización frente a los de educación (M=3.30; M=2.14), y sobre todo con respecto a los de economía (M=2.40; M=1.37). Con respecto a la observación de las simulaciones, también existen diferencias significativas entre comunicación (M=2.56) y educación (M=2.57) frente a economía (M=1.56).

En la encuestación se dan diferencias en dos de los cuatro medios que conforman la dimensión. En el caso de las entrevistas orales, aunque en general su uso es muy bajo (M=1.35), es el profesorado de economía y empresa (M=0.96) el que manifiesta un menor uso. También se han encontrado diferencias significativas en el uso de los grupos de discusión o grupos focales, siendo el profesorado de comunicación (M=2.50) el que expresa una mayor utilización, frente al de educación (M=2.13) y el de economía y empresa (M=1.59).

En el uso de objetos digitales o presentaciones multimedia, las diferencias se encuentran entre el profesorado de educación (M=3.45) y comunicación (M=3.33), frente al de economía y empresa (M=2.74). Sin embargo, en el uso de las pruebas objetivas, se invierte la situación, siendo el de economía y empresa (M=2.53) el que manifiesta una mayor utilización que el de comunicación (M=2.22) o el de educación (M=1.81). Algo similar sucede con el uso de las pruebas escritas con preguntas que exigen respuesta breve dado que la puntuación de economía y empresa es la más alta (M=2.02) seguida de educación (M=1.38) y comunicación (M=1.33). La utilización de pruebas escritas de mapas conceptuales está poco extendida, aunque en este caso su uso es mayor en comunicación (M=2.00) que en educación (M=1.09) y en economía y empresa (M=0.68). Los diarios también son poco utilizados y al igual que en el caso anterior, hay un mayor uso en comunicación (M=2.06) que en educación (M=1.19) y en economía y empresa (M=0.53). La evaluación de los proyectos está, en general, más extendida y las diferencias se encuentran entre comunicación (M=4.00), con respecto a educación (M=3.42) y economía y empresa (M=2.70). Por último, el uso del portafolio no está muy extendido (M=1.55), aunque se expresa un uso mayor por parte del profesorado de educación (M=2.10), frente al de comunicación (M=1.89) y el de economía y empresa (M=0.76).

Al analizar las diferencias en los instrumentos se expresa una mayor utilización de las escalas de estimación por el profesorado de comunicación (M=1.78) y educación (M=1.56), frente a economía y empresa (M=0.89). Sobresale el uso manifestado de las rúbricas por el profesorado de educación (M=3.56) frente al de economía y empresa (M=2.31) y el de comunicación (M=1.78). Se manifiesta un mayor uso

del argumentario valorativo por parte del profesorado de comunicación (M=3.17) y educación (M=2.90) frente al de economía y empresa (M=2.23).

Diferencias en función de la seguridad y satisfacción con la evaluación

En 10 de los 23 ítems se presentan diferencias significativas. En la observación en cuatro de los seis medios considerados; en la encuestación en tres de los cuatro medios y, por último, en dos de los nueve medios en los que se hace uso del análisis de documentos y artefactos. En los instrumentos de evaluación solo se han encontrado diferencias en el uso del argumentario evaluativo.

Al comparar los cuatro niveles, tanto en la observación de la participación activa, como en las prácticas del alumnado, las pruebas de resolución de problemas o las pruebas de desempeño, los docentes situados en el nivel 4 (mayor seguridad evaluadora) expresaron un mayor grado de utilización frente a los restantes niveles.

En el caso de las entrevistas orales, el profesorado del nivel 4 (M=1.77) expresa una mayor utilización que el profesorado del nivel 1 (M=1.06), nivel 2 (M=1.35) o nivel 3 (M=0.97). La utilización de los grupos de discusión o grupos focales también está más extendida entre el profesorado del nivel 4 (M=2.24) que el del nivel 1 (M=1.67). Se expresa un mayor uso de las pruebas de resolución de problemas por los docentes de los niveles 3 (M=3.43) y 4 (M=3.30) frente a los niveles 2 (M=2.88) y 1 (M=2.70).

En relación al uso de los objetos digitales o presentaciones multimedia, una vez más los docentes con una seguridad evaluadora mayor (nivel 4) expresan mayor asiduidad en su uso (M=3.36) que los del nivel 1 (M=2.84). De la misma forma, el profesorado del nivel 4 hace un mayor uso de las pruebas objetivas (M=2.84) que el resto de los niveles.

Por último, en los instrumentos de evaluación solo se han dado diferencias significativas en la utilización del argumentario valorativo. También esta vez, el profesorado del nivel 4 expresa un mayor grado de uso (M=3.04) que el resto del profesorado del nivel 1 (M=2.49), nivel 2 (M=2.54) o nivel 3 (M=2.47).

Diferencias en función del género y la experiencia docente

Sobre la base del género, tan solo se han encontrado diferencias en el uso de dos medios de evaluación y en uno de los instrumentos. En la observación de la participación de los estudiantes, las mujeres expresan un mayor grado de utilización (M=3.91) que los hombres (M=3.51) o el colectivo otros (M=2.00). Igualmente, se manifiesta por parte de las mujeres un mayor uso (M=3.42) del análisis de objetos digitales o presentaciones multimedia frente a los hombres (M=2.84) y otros

(M=3.00). Por último, el uso de las rúbricas es manifestado en mayor medida por las mujeres (M=3.13) en comparación con los hombres (M=2.58) o el nulo uso por parte de otros.

En referencia a la experiencia docente solo se han encontrado diferencias en el uso de tres de los medios de evaluación presentados (simulaciones, objetos digitales o presentaciones multimedia y proyectos) y en la utilización de rúbricas. Prácticamente en estos cuatro casos, a medida que aumentan los años de experiencia baja la puntuación obtenida en esos ítems. Así, por ejemplo, en el caso de las rúbricas, el profesorado más novel (M=3.54) manifiesta utilizar las mismas en una mayor proporción que el profesorado con una experiencia entre 11 y 20 años (M=2.93) y sobre todo con mayor diferencia sobre el profesorado más experimentado (M=2.51).

DISCUSIÓN Y CONCLUSIONES

En este estudio se pretendía comprobar, en primer lugar, cuáles son los medios e instrumentos de evaluación utilizados en los másteres de Ciencias Sociales. En segundo lugar, analizar si existen diferencias en función de ciertas variables contextuales y personales. Y la tercera cuestión de investigación estaba centrada en si se percibe alguna evolución en la utilización de medios e instrumentos de evaluación.

Como se ha manifestado, la validez, pertinencia y adecuación de los medios e instrumentos utilizados en la evaluación permitirán realizar inferencias justas y fundamentadas sobre los avances del estudiantado (Sadler, 2016). En estudios previos se ha comprobado que en los títulos de grado en distintas universidades españolas, aunque ha habido un ligero cambio, sigue prevaleciendo el examen final como medio para la evaluación. Esto ha sido corroborado tanto en estudios realizados analizando las guías docentes de las asignaturas (Ibarra-Sáiz & Rodríguez-Gómez, 2010; Rodríguez-Gómez et al., 2013; Panadero et al., 2019) como en estudios basados en la opinión de los estudiantes (Lukas et al., 2011; Lukas et al., 2016).

Otros estudios evidencian la relación entre la calidad de la evaluación y el desarrollo competencial del estudiantado (Ibarra-Sáiz et al., 2020a) y, como indica Boud (2020), el uso de medios de evaluación inadecuados puede conducir a un aprendizaje deficiente, por lo que es necesaria la utilización de medios que produzcan aprendizajes de alto nivel en los estudiantes. En este sentido, los resultados obtenidos muestran algunas diferencias en los medios e instrumentos de evaluación con respecto a los estudios previos de grado. De esta forma, se ha comprobado que los docentes utilizan la observación para analizar la participación, resolución de problemas, prácticas del alumnado o pruebas de desempeño. Mucho

menor es su uso en el análisis de simulaciones y talleres, que no son tan habituales en Ciencias Sociales.

En general la encuestación es menos utilizada por el profesorado. Es algo más habitual su uso en las pruebas de resolución de problemas o de desempeño, y apenas se utilizan los grupos focales o las entrevistas orales.

Respecto al análisis de documentos y artefactos sobresalen el análisis de los objetos digitales o presentaciones multimedia y el análisis de proyectos o de pruebas de resolución de problemas. Un tercio del profesorado señala que utiliza las pruebas objetivas, pruebas escritas de respuesta breve o pruebas escritas de desarrollo tipo ensayo. Esta es una gran diferencia respecto a su uso en los grados como ha quedado demostrado en estudios previos. Existe un uso casi residual del portafolio, diarios y pruebas escritas de mapas conceptuales.

Por último, los instrumentos de evaluación más utilizados por el profesorado son el argumentario valorativo y, sobre todo, las rúbricas, cuyo uso se ha extendido estos últimos años y es utilizado por más de la mitad de los docentes. Las listas de control y las escalas de estimación tienen todavía un uso reducido.

Las diferencias reseñadas con respecto a estudios anteriores pueden ser debidas a distintas causas. Una de ellas puede ser el número de estudiantes, que es muy diferente en los estudios de grado (mucho mayor habitualmente) frente a los másteres (mucho menor por lo general). Otra causa puede ser la cercanía de los másteres con el futuro profesional de los estudiantes, lo que exige a los docentes utilizar medios de evaluación más cercanos a la evaluación auténtica. Es decir, se proponen el uso de tareas y procesos de evaluación representativos de los tipos de tareas y procesos que se pueden encontrar en la práctica profesional. Estos resultados corroboran lo señalado por Boud (2020) cuando afirma que se ha pasado del uso de las clásicas pruebas, test o exámenes finales centrados en la reproducción del conocimiento, a un conjunto de medios de evaluación a través de los cuales se pretende integrar y dar coherencia al aprendizaje. Los resultados obtenidos, aunque todavía lejos de una situación ideal, suponen un ligero cambio en la diversidad de los medios e instrumentos de evaluación dado que lentamente se va centrando el foco en las ejecuciones del estudiante, en lo que produce y hace, fomentando el pensamiento crítico y la creatividad (Boud, 2020), y abre las puertas a la mejora en los diseños de las tareas de evaluación (Ibarra-Sáiz et al., 2021) con niveles más altos de calidad.

La tercera cuestión del estudio se centraba en comprobar si se podrían asociar ciertas diferencias en el uso de medios e instrumentos de evaluación con determinadas variables contextuales. En este sentido, los resultados evidencian que tanto la universidad como el ámbito de conocimiento son las variables en las que se han encontrado más diferencias significativas. En ambos casos, estas diferencias se han dado en el 65% de los ítems analizados. Un tercer elemento

diferenciador lo constituye el nivel de satisfacción y seguridad que tiene el profesorado con el sistema de evaluación que practica, ya que en el 48% de los ítems se presentan diferencias entre los diferentes niveles. Las diferencias sobre la base de la experiencia profesional se evidencian en un 17% de los ítems, y en un 13% en el caso del género. Este hecho pone de manifiesto la relevancia de las políticas de desarrollo profesional docente a nivel de universidad desde una perspectiva estratégica, así como considerar las diferencias en la cultura docente, construidas a partir del ámbito de conocimiento del profesorado, para fundamentar cambios en las prácticas evaluativas coherentes con sus contextos inmediatos, y basados en la necesaria actitud reflexiva crítica.

Una limitación de este estudio viene dada por la muestra productora de los datos, ya que no fue seleccionada aleatoriamente y se circunscribe a seis universidades públicas. No obstante, es preciso matizar, en primer lugar, que cada una de las universidades participantes pertenecen a comunidades autónomas diferentes y con agencias de calidad independientes; además, las características demográficas del profesorado participante presentan suficiente heterogeneidad, por lo que se sustenta suficientemente la representatividad muestral.

Investigaciones futuras podrían ampliar la muestra a universidades públicas y privadas, a otras ramas de conocimiento y a los títulos de grado, lo que permitiría ampliar la generalización de los resultados. En cualquier caso, hay una necesidad evidente de estudios que profundicen, no solo en la tipología de los medios e instrumentos de evaluación que utiliza el profesorado en su práctica evaluativa sino, sobre todo, en qué características edumétricas (Dochy, 2009) tienen estos medios e instrumentos, y cómo son utilizados por profesorado y estudiantado, en tanto que constituyen elementos que modulan el aprendizaje de los estudiantes. Desde esta perspectiva, sería conveniente analizar estos medios e instrumentos de evaluación bajo los criterios edumétricos de la complejidad cognitiva, autenticidad de las tareas, equidad y justicia, transparencia de los procedimientos de evaluación y la influencia de la evaluación en la educación, lo que supone analizar la validez de las tareas de evaluación, la validez de las puntuaciones en las evaluaciones del desempeño, la generalizabilidad de la evaluación y su validez consecuente. Esta línea de investigación podrá enfocarse metodológicamente desde una perspectiva cualitativa, mediante estudios de caso que profundicen en la comprensión de la práctica evaluativa en educación superior, y habrá que considerar como marco de referencia conceptual las aportaciones y características del enfoque evaluativo de la evaluación como aprendizaje (Boud, 2022; Yan & Boud, 2022).

En este estudio se ha evidenciado cómo el profesorado universitario percibe su práctica evaluativa, desde la consideración de la diversidad de los medios e instrumentos de evaluación que manifiestan utilizar. Además, la aportación del autoinforme RAPEVA facilitará la adaptación o réplica de este estudio en otros contextos que se han sugerido como posibles vías futuras de investigación.

Las diferencias encontradas a partir de la universidad de procedencia, el ámbito de conocimiento en el que se trabaja o la satisfacción manifestada con el sistema de evaluación hacen necesaria la revisión crítica y el impulso de políticas de formación y desarrollo profesional docente con carácter general, y de forma específica en lo que a los enfoques de la evaluación se refiere. Solo en la medida en que las universidades propicien un mayor desarrollo profesional y procesos formativos críticos, sustentados en los avances en el conocimiento alcanzado sobre la evaluación en educación superior en las últimas décadas (Ibarra-Sáiz et al., 2020b), se podrán ir integrando en la práctica evaluativa medios e instrumentos que favorezcan y potencien el aprendizaje de alto nivel del estudiantado.

AGRADECIMIENTOS

Este trabajo ha sido posible gracias al Proyecto FLOASS – *Resultados y analíticas de aprendizaje en la educación superior: Un marco de acción desde la educación sostenible* (Ref. RTI2018-093630-B100), financiado por el Ministerio de Ciencia, Innovación y Universidades en el Programa Estatal de I+D+i Orientada a los Retos de la Sociedad, la Agencia Estatal de Investigación, el Fondo Europeo de Desarrollo Regional, y el apoyo de la Cátedra UNESCO en *Evaluación, Innovación y Excelencia en Educación*.

REFERENCIAS BIBLIOGRÁFICAS

Anderson, M.J. (2017). Permutational Multivariate Analysis of Variance (PERMANOVA). *Wiley StatsRef: Statistics Reference Online*, 1–15. https://doi.org/10.1002/9781118445112.stat07841

Biggs, J. (14-15 de mayo de 2015). *Assessment in a constructively system*. [Ponencia de congreso]. International Conference Assessment for Learning in Higher Education 2015, Hong Kong, China.

Biggs, J., & Tang, C. (2011). *Teaching for quality learning at university. What the students does* (4th ed.). McGraw-Hill-SRHE & Open University Press.

Boud, D. (2020). Challenges in reforming higher education assessment: a perspective from afar. *RELIEVE*, *26*(1), Artículo M3. https://doi.org/10.7203/relieve.26.1.17088

Boud, D. (2022). Assessment-as-learning for the development of students' evaluative judgement. En Z. Yan, & L. Yang (Eds), *Assessment as learning. Maximising opportunities for student learning and achievement* (pp. 25–37). Routledge.

Brown, S., & Pickford, R. (2013). *Evaluación de habilidades y competencias en Educación Superior*. Narcea.

Creswell, J. W. (2015). *A concise introduction to mixed methods research*. SAGE Publications.

Dochy, F. (2009). The edumetric quality of new modes of assessment: Some issues and prospect. En G. Joughin (Ed.), *Assessment, learning and judgement in higher education* (pp. 85–114). Springer Science & Business Media B.V.

European Centre for Development of Vocational Training. (2014). *Terminology of European education and training policy. A selection of 130 key terms.* Publications Office of the European Union.

Hair, J. F., Hult, G. T. M., Ringle, C. M., & Sarstedt, M. (2022). *A primer on partial least squares structural equation modeling (PLS-SEM)* (3rd ed.). SAGE Publications.

Henseler, J. (2021). *Composite-based structural equation modeling. Analyzing latent and emergent variables.* Guilford Press.

Hwang, H., & Takane, Y. (2015). *Generalized structured component analysis: A component-based approach to structural equation modeling*. CRC Press.

Ibarra-Sáiz, M.S., & Rodríguez-Gómez, G. (19-21 de junio de 2019). *FLOASS - Learning outcomes and learning analytics in higher education: An action framework from sustainable assessment*. [Póster]. XIX Congreso Internacional de Investigación Educativa. Investigación comprometida para la transformación social, Madrid, España.

Ibarra-Sáiz, M.S., & Rodríguez-Gómez, G. (2010). Aproximación al discurso dominante sobre la evaluación del aprendizaje en la universidad. *Revista de Educación*, (351), 385–407.

Ibarra Saiz, M. S., & Rodríguez Gómez, G. (2014). Modalidades participativas de evaluación: Un análisis de la percepción del profesorado y de los estudiantes universitarios. *Revista de Investigación Educativa, 32*(2), 339-361. https://dx.doi.org/10.6018/rie.32.2.172941

Ibarra-Sáiz, M.S., & Rodríguez-Gómez, G. (2020). Evaluando la evaluación. Validación mediante PLS-SEM de la escala ATAE para el análisis de tareas de evaluación. *RELIEVE*, *26*(1), Artículo M4. https://doi.org/10.7203/relieve.26.1.17403

Ibarra-Sáiz, M.S., Rodríguez-Gómez, G., & Boud, D. (2020a). Developing student competence through peer assessment: the role of feedback, self-regulation and evaluative judgement. *Higher Education*, *80*(1), 137–156. https://doi.org/10.1007/s10734-019-00469-2

Ibarra-Sáiz, M.S., Rodríguez-Gómez, G., Boud, D., Rotsaert, T., Brown, S., Salinas Salazar, M. L., & Rodríguez Gómez, H. M. (2020b). El futuro de la evaluación en educación superior. *RELIEVE, 26*(1), Artículo M1. https://doi.org/10.7203/relieve.26.1.17323

Ibarra-Sáiz, M.S., Rodríguez-Gómez, G., & Boud, D. (2021). The quality of assessment tasks as a determinant of learning. *Assessment & Evaluation in Higher Education*, *46*(6), 943–955. https://doi.org/10.1080/02602938.2020.1828268

JASP Team. (2022). *JASP (Version 0.16.1)*. https://jasp-stats.org/

Johnson, R. L., & Morgan, G. B. (2016). *Survey scales. A guide to development, analysis, and reporting*. The Guilford Press.

Lukas, J. F., Santiago, K., & Murua, H. (2011). Unibertsitateko ikasleen ikaskuntzara bideratutako ebaluazioa. *Tantak, 23*(1), 77-97.

Lukas, J.F., Santiago, K., Lizasoain, L., & Etxeberria, J. (2017). Percepciones del alumnado universitario sobre la evaluación. *Bordón, 69*(1), 103-122. https://doi.org/10.13042/Bordon.2016.43843

Mateo Andrés, J., & Martínez Olmo, F. (2008). *La evaluación alternativa de los aprendizajes.* Octaedro.

Panadero, E., Fraile, J., Fernández Ruiz, J., Castilla-Estévez, D., & Ruiz, M. A. (2019). Spanish university assessment practices: examination tradition with diversity by faculty. *Assessment & Evaluation in Higher Education*, *44*(3), 379–397. https://doi.org/10.1080/02602938.2018.1512553

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.r-project.org/

Rodríguez-Gómez, G., & Ibarra-Sáiz, M. S. (Eds.). (2011). *e-Evaluación orientada al e-Aprendizaje estratégico en Educación Superior*. Narcea.

Rodríguez-Gómez, G., Ibarra-Sáiz, M. S., & García-Jiménez, E. (2013). Auto-evaluación, evaluación entre iguales y coevaluación: conceptualización y práctica en las universidades españolas. *Revista de Investigacion en Educación*, *11*(2), 198–210.

Sadler, D.R. (2016). Three in-course assessment reforms to improve higher education learning outcomes. *Assessment & Evaluation in Higher Education*, *41*(7), 1081–1099. https://doi.org/10.1080/02602938.2015.1064858

Yan, Z., & Boud, D. (2022). Conceptualising assessment-as-learning. En Z. Yan & L. Yang (Eds.), *Assessment as learning. Maximising opportunities for student learning and achievement* (pp. 11–24). Routledge.