



**ISSN: 1989-0397**

# Revista Iberoamericana de Evaluación Educativa

**Noviembre 2020 - Volumen 13, número 2**

**doi: 10.15366/rie2020.13.2**

**[rinace.net/rie/](http://rinace.net/rie/)  
[revistas.uam/rie](http://revistas.uam/rie)**

**UAM** Universidad Autónoma  
de Madrid



Red Iberoamericana  
de Investigación  
sobre Cambio  
y Eficacia Escolar

# CONSEJO EDITORIAL

## DIRECTOR

F. Javier Murillo

## EDITORA

Nina Hidalgo Farran

## CONSEJO DIRECTIVO

Marcela Gajardo (Programa de Promoción de la Reforma Educativas de América Latina y El Caribe, PREAL)

Sergio Martinic (Pontificia Universidad Católica de Chile)

Carlos Pardo (Instituto Colombiano para la Evaluación de la Educación, ICFES)

Margarita Poggi (Instituto Internacional de Planeamiento de la Educación -IIPE-. UNESCO, Argentina)

Francisco Soares (Universidade Federal de Minas Gerais, Brasil)

## CONSEJO CIENTÍFICO

Juan Manuel Álvarez. Universidad Complutense de Madrid, España.

Patricia Arregui. Grupo de Análisis para el Desarrollo (GRADE), Perú.

Daniel Bogoya. Universidad Pedagógica Nacional, Colombia.

Nigel Brooke. Universidade Federal de Minas Gerais, Brasil.

Leonor Cariola. Ministerio de Educación, Chile.

María do Carmo Clímaco. Universidade Lusófona de Humanidades e Tecnologias (ULHT), Portugal.

Cristian Cox. Pontificia Universidad Católica de Chile.

Santiago Cueto. Grupo de Análisis para el Desarrollo (GRADE).

Tabaré Fernández. Universidad de la República, Uruguay.

Juan Enrique Froemel. Universidad UNIACC, Chile.

Rubén Klein. Fundação Cesgranrio, Brasil.

Luis Lizasoain. Universidad del País Vasco/Euskal Herriko Unibertsitatea, España.

Jorge Manzi. MIDE-UC, Pontificia Universidad Católica de Chile.

Joan Mateo. Universidad de Barcelona, España.

Liliana Miranda. Ministerio de Educación de Perú.

Margarita Peña. Instituto Colombiano para la Evaluación de la Educación, ICFES.

Dagmar Raczynski. Asesorías para el Desarrollo, Chile.

Héctor Rizo. Universidad Autónoma de Occidente, Colombia.

Mario Rueda. Universidad Nacional Autónoma de México.

Guadalupe Ruíz. Universidad Autónoma de Aguascalientes, México.

Ernesto Schiefelbein. Universidad Autónoma de Santiago, Chile.

Alejandra Schullmeyer. Instituto Nacional de Estudios Pedagógicos, Brasil.

Javier Tejedor. Universidad de Salamanca, España.

Flavia Terigi. Universidad de Buenos Aires, Argentina.

Alexander Ventura. Universidade de Aveiro, Portugal.

# ÍNDICE

## Temática libre

- Métodos de Avaliação das Aprendizagens em Universidades Públicas Portuguesas: Um Estudo Quantitativo** 5

*Eva Lopes Fernandes, Maria Assunção Flores, Irene Cadime y Clara Pereira Coutinho*

- Evaluación basada en los Resultados de Aprendizaje: Una Experiencia en la Universidad** 27

*Eugenio Astigarraga Echeverría, Arantza Mongelos García y Xavier Carrera Farran*

- Diferencias de Género y Estudios de Acceso en las Creencias del Alumnado de Grado en Educación Infantil sobre el Desarrollo de la Autonomía en el Ciclo 0-3** 49

*Elena Herrán Izagirre, Nuria Galende Pérez y Gorka Etxebarria Elordui*

- Evaluación Auténtica y Evaluación Orientada al Aprendizaje en Educación Superior. Una Revisión en Bases de Datos Internacionales** 67

*Emilio José Barrientos-Hernán, Víctor M. López-Pastor y Darío Pérez-Brunicardi*

- La evaluación para el aprendizaje en la educación superior: retos de la alfabetización del profesorado** 85

*Rafael López-Meseguer y Manuel T. Valdés*

- El Examen de Ingreso a la Universidad Nacional Autónoma de México: Evidencias de Validez de una Prueba de Alto Impacto y Gran Escala** 107

*Melchor Sánchez Mendiola, Manuel García Minjares, Adrián Martínez González y Enrique Buzo Casanova*



# Métodos de Avaliação das Aprendizagens em Universidades Públicas Portuguesas: Um Estudo Quantitativo

## Assessment Methods in Portuguese Public Universities: A Quantitative Study

Eva Lopes Fernandes \*  
Maria Assunção Flores  
Irene Cadime  
Clara Pereira Coutinho

Universidade do Minho, Portugal

Este artigo enquadra-se num projeto de investigação mais amplo que procura identificar os métodos de avaliação mais valorizados e mais utilizados por professores universitários no contexto português. Os resultados foram recolhidos através de um inquérito por questionário com professores de cinco universidades públicas portuguesas (n=185), de todas as categorias profissionais e ciclos de estudos, das seguintes áreas científicas: Ciências Exatas, Ciências da Engenharia e da Tecnologia, Ciências Médicas e da Saúde, Ciências Sociais e Humanidades. A análise fatorial exploratória indicou uma organização das escalas sobre a valorização e frequência de utilização de métodos de avaliação em três fatores: métodos coletivos, portefólios e reflexões, e métodos individuais, tendo-se procedido ao tratamento dos testes e exames escritos como variável observável. Os resultados das estatísticas descritivas identificam uma maior valorização e frequência de utilização de testes e exames escritos e menor valorização e frequência de utilização de portefólios e reflexões. Os dados sugerem ainda diferenças estatisticamente significativas na utilização de métodos de avaliação em função dos ciclos de estudos, área de conhecimento e formação pedagógica.

**Palavras-Chave:** Ensino superior; Métodos de avaliação; Avaliação das aprendizagens; Professores universitários.

This article is part of a broader research project aimed at identifying the most valued and most used assessment methods by faculty in the Portuguese context. The results were collected through a survey with faculty from five Portuguese public universities (n=185) in all professional categories and cycles of study, of the following scientific areas: Hard Sciences, Engineering and Technology Sciences, Medical and Health Sciences, Social Sciences and Humanities. The exploratory factor analysis (EFA) results indicated a three factors model for both scales: collective methods, portfolios and reflections, and individual methods. The item "written tests and exams" was treated in subsequent analysis independently as observable variable. The results of descriptive statistics identify a greater valorisation and use of written tests and exams and a lower valorisation and use of portfolios and reflections. Data also suggest statistically significant differences in the use of assessment methods according to the cycles of study, the area of knowledge and pedagogical training.

**Keywords:** Higher education; Assessment methods; Assessment of learning; Faculty.

---

\*Contacto: evalopesfernandes@gmail.com

## 1. Introdução

A avaliação no contexto do Ensino Superior constitui uma temática que tem recebido uma atenção crescente por parte dos decisores políticos, dos investigadores, dos gestores, dos professores e outros *stakeholders* na medida em que influencia a forma como os estudantes organizam o seu tempo e mobilizam os seus esforços (Fernandes, 2015; Myers e Myers, 2015) e as suas perceções sobre o modo como as aprendizagens se desenvolvem (Brown e Knight, 1994). Daí a importância da seleção dos métodos de avaliação mais adequados em função dos objetivos de ensino e aprendizagem (Pereira e Flores, 2016).

Estudos realizados no ensino superior demonstram que os instrumentos de avaliação tradicionais são os mais utilizados (principalmente o exame escrito) (Barreira et al., 2015; Pereira, 2016; Pereira e Flores, 2016), destacando-se o sistema de classificação e a hierarquização (Pereira e Flores, 2013; Perrenoud, 1999). Outras perspetivas de avaliação, como, por exemplo, o *assessment for learning* (McDowell et al., 2011) apontam para um papel de cada vez maior responsabilização dos estudantes no processo de avaliação e de aprendizagem. Este alargamento de papéis aos diversos atores é extensível aos professores, que necessitam de assumir o processo de ensino e aprendizagem de forma mais autónoma, colaborativa e integrada, através de projetos partilhados; da produção coletiva do conhecimento; no sentido de tornar o processo de aprendizagem mais criativo (Flores e Veiga Simão, 2007), compreendendo a avaliação como uma parte integrante da aprendizagem (Zabalza, 2008).

Neste artigo são apresentados dados decorrentes de um projeto mais vasto sobre avaliação no Ensino Superior na perspetiva de docentes universitários, nomeadamente no que diz respeito à valorização e utilização de métodos de avaliação.

A literatura existente distingue métodos tradicionais de métodos alternativos de avaliação (Duncan e Buskirk-Cohen, 2011), destacando-se práticas centradas no professor e centradas no aluno (Burkšaitienė e Teresevičienė, 2008; Fernandes, 2015; Myers e Myers, 2015), e reconhecendo-se, no âmbito destas últimas, o espaço privilegiado da sala de aula na organização de ambientes inovadores e facilitadores da aprendizagem (Black e William, 1998) e no desenvolvimento de formas inovadoras de estruturar o ensino e a avaliação (Fernandes, 2015). A transformação das práticas pedagógicas encerra uma mudança no papel das instituições do ensino superior enquanto contextos que existem para produzir conhecimento (Burkšaitienė e Teresevičienė, 2008). Esta visão assume especial importância no contexto do Processo de Bolonha (Flores e Veiga Simão, 2007), desafiando os professores a promoverem oportunidades de aprendizagem mais questionadoras, inovadoras e criativas (Fernandes e Flores, 2012; Zabalza, 2008).

Os métodos tradicionais, comumente usados no contexto do Ensino Superior (Duncan e Buskirk-Cohen, 2011; Pereira e Flores, 2016), sobretudo no contexto de turmas grandes, a par da sobrecarga de trabalho dos professores (Myers e Myers, 2015), podem enfatizar a reprodução e memorização (Perrenoud, 1999) e baixos níveis de compreensão (Dochy, et al, 2007). Os métodos tradicionais (nomeadamente o teste ou exame escrito) ocorrem num momento pré-determinado, centram-se no produto ou resultados e apresentam usualmente uma forte componente individual (Hadji, 1994), representando um meio de perceber o que se passa na sala de aula (Phelps, 2017).

Por seu turno, os chamados métodos de avaliação centrados no aluno (Myers e Myers, 2015; Webber e Tschepikow, 2013) permitem o desenvolvimento de competências

técnicas e transversais, por exemplo, de resolução de problemas, fomentando um maior envolvimento dos alunos no processo de aprendizagem (Myers e Myers, 2015). Normalmente, estes métodos implicam tarefas mais globais, desenvolvidas ao longo do tempo e que podem incluir, simultaneamente, o processo e o produto e a dimensão individual e coletiva, estimulando a autonomia, a colaboração, responsabilidade, o *feedback* construtivo, a interação com os pares e a construção de conhecimento (Pereira e Flores, 2013; Webber e Tschepikow, 2013), o desenvolvimento de competências e o aprofundamento das aprendizagens (Brew, Riley e Walta, 2009). Estes métodos incluem, por exemplo, os trabalhos práticos experimentais, o trabalho de projeto ou as reflexões (Struyven, Dochy e Steven, 2005; Webber e Tschepikow, 2013). O portefólio destaca-se, pela sua especificidade, ao apresentar um carácter interativo, dinâmico, aberto e global. Trata-se de um método peculiar que encerra diferentes funções e tarefas, e que pode ser individual ou coletivo. Os métodos alternativos são usados ao longo do tempo, potenciando a revisão e análise dos elementos de avaliação produzidos e reconhecendo o papel essencial do *feedback* no processo de aprendizagem do aluno e na sua autorregulação (Carless et al., 2011). Apesar dos seus reconhecidos benefícios, são-lhes, igualmente, apontadas algumas limitações ao nível da prestação de contas e da dificuldade em usar a informação produzida do ponto de vista sumativo (Maclellan, 2004), o que, atendendo aos inúmeros desafios no campo da avaliação, pode justificar o seu uso limitado no contexto do ensino superior.

Estudos recentes apontam para a necessidade de mais investigação neste domínio, nomeadamente no que se refere aos métodos de avaliação mais utilizados e ao seu impacto na aprendizagem dos alunos (Watering et al., 2008), à comparação entre práticas de avaliação em diferentes áreas, instituições e países (Gilles, Detroz e Blais, 2010), à eficácia dos chamados métodos alternativos de avaliação ou centrados nos estudantes (Segers, Gijbels e Thurlings, 2008) e à necessidade de analisar as práticas de avaliação em articulação com mecanismos de *feedback* (Flores et al., 2014).

No sentido de responder a essa necessidade, este estudo pretende conhecer as conceções e práticas de avaliação de docentes do Ensino Superior após a implementação do Processo de Bolonha, mais concretamente através da identificação e análise dos métodos de avaliação mais valorizados e utilizados. Para o efeito, procedeu-se à validação de um instrumento de recolha de dados sobre valorização e utilização de métodos de avaliação adaptado de Pereira (2011, 2016) e Gonçalves (2016), que foi aplicado em cinco universidades públicas portuguesas. O instrumento foi concebido com o objetivo de alargar o espectro do estudo original de Pereira (2016) às perspetivas e práticas de professores universitários no sentido de dar respostas às seguintes questões de investigação:

- 1) Que importância atribuem aos métodos de avaliação?
- 2) Qual a frequência da utilização dos diferentes métodos de avaliação?
- 3) Quais as variáveis associadas à valorização e utilização de diferentes métodos de avaliação?
- 4) Qual a relação entre a valorização e a utilização de diferentes métodos de avaliação?

## 2. Métodos

Este artigo enquadra-se num projeto de investigação mais amplo realizado no âmbito do Doutoramento em Ciências da Educação (Ref. SFRH/BD/103291/2014) que, por sua vez, se integra um projeto de investigação mais alargado, intitulado “Assessment in Higher Education: The potential of alternative methods” (Ref. PTDC/MHC-CED/2703/2014), projetos financiados pela Fundação Nacional para a Ciência e Tecnologia (FCT). Este estudo teve como objetivos:

- 1) Analisar a influência das variáveis demográficas na valorização e utilização de métodos de avaliação;
- 2) Analisar a influência das variáveis profissionais na valorização e utilização de métodos de avaliação;
- 3) Analisar a influência das variáveis relacionadas com as práticas de avaliação e com o Processo de Bolonha na valorização e utilização de métodos de avaliação;
- 4) Testar a existência de uma correlação entre a valorização e a utilização de métodos de avaliação.

### 1.1. Participantes

Os resultados foram recolhidos através de um inquérito por questionário, aplicado entre fevereiro e julho de 2017, em cinco universidades públicas portuguesas. A amostra é composta por 185 professores universitários de todas as categorias profissionais e ciclos de estudos. Tendo em conta os objetivos de investigação optou-se por uma amostra de conveniência (Coutinho, 2014), definida de acordo com as áreas científicas identificadas pela FCT, assegurando que as cinco universidades envolvidas tivessem uma oferta educacional semelhante nas cinco áreas selecionadas.

No quadro 1 apresenta-se uma descrição detalhada das características dos participantes: são principalmente do sexo feminino, têm mais de 45 anos, são professores associados ou auxiliares, com mais de 15 anos de experiência e com formação pedagógica. Lecionam nas seguintes áreas de conhecimento: Ciências Médicas e da Saúde, Ciências Exatas, Ciências da Engenharia e Tecnologia, Ciências Sociais e Humanidades e em diferentes cursos: 83,8% lecionam nos cursos de graduação; 77,3% em cursos de mestrado; 41,5% em cursos de Mestrado Integrado; 55,8% em cursos de doutoramento; e, 1,7% em outros cursos (por exemplo, cursos não conferentes de grau).

### 1.2. Instrumento de recolha de dados

O questionário utilizado neste estudo é uma versão adaptada dos “Métodos de Avaliação no Ensino Superior”, utilizado no estudo de Pereira (2011, 2016) e da secção sobre utilização e importância dos métodos de avaliação do “Inventário de Conceções de Avaliação” (ICA) de Gonçalves (2016), centrando-se em dois aspetos fundamentais: 1) importância atribuída aos diferentes métodos de avaliação; e, 2) métodos os métodos de avaliação mais utilizados pelos docentes universitários. Ambas as escalas são constituídas pelos mesmos 14 itens (consultar quadro 2). Para cada item da primeira escala, os participantes classificaram a importância atribuída ao método utilizando uma escala *likert* de quatro pontos que varia entre 1 (nada importante) e 4 (muito importante). Era ainda oferecida a possibilidade de os participantes selecionarem a opção “Não se aplica”. Estas respostas foram codificadas com o valor mais baixo da escala de resposta (1). Na segunda

escala, os participantes indicaram a frequência da utilização de cada método, recorrendo a uma escala de *likert* de quatro pontos entre 1 (nada utilizados) e 4 (muito utilizados)

Quadro 1. Caracterização dos participantes

<b>CARACTERÍSTICAS DEMOGRÁFICAS</b>		
<b>Universidade</b>	<b>N</b>	<b>%</b>
A	36	19,5
B	34	18,4
C	60	32,4
D	36	19,5
E	19	10,2
<b>Sexo</b>		
Masculino	74	40,0
Feminino	87	47,0
Sem informação	24	13,0
<b>Idade</b>		
Menos de 45 anos de idade	82	44,3
Mais de 45 anos de idade	103	55,7
<b>Área de Conhecimento</b>		
Ciências Médicas e da saúde	21	11,4
Ciências Exatas	16	8,6
Ciências da Engenharia e Tecnologia	50	27,0
Ciências sociais	77	41,6
Humanidades	21	11,4
<b>Categoria Profissional</b>		
Professor Catedrático	10	5,4
Professor Associado/Auxiliar com Agregação	19	10,3
Professor Associado/Auxiliar	132	71,3
Outros	24	13,0
<b>Experiência Docente</b>		
Menos de 15 anos	54	29,2
Mais de 15 anos	131	70,8
<b>Formação Pedagógica</b>		
Sim	117	63,2
Não	63	34,1
Sem informação	5	2,7

Fonte: Elaboração própria.

### 1.3. Procedimentos

No contacto com as instituições e ao longo de toda a investigação foram respeitadas as questões éticas de investigação na área das Ciências Sociais e Humanas reconhecidas internacionalmente, tendo o estudo sido aprovado pela Comissão de Ética da Universidade do Minho (Ref. SECSH035/2016 e SECSH036/2016). Foram contactados todos os autores dos instrumentos de recolha de dados, tendo-se obtido autorização para o seu uso e adaptação ao contexto português. Foram celebrados protocolos de investigação com as faculdades e escolas/institutos envolvidos. Todos os participantes assinaram um consentimento informado voluntário após uma breve explicação do projeto e dos objetivos de investigação.

### 1.4. Análise estatística

As análises estatísticas foram realizadas com recurso ao software *IBM SPSS Statistics 24*. Por forma a explorar a estrutura interna do instrumento recorreu-se à análise fatorial exploratória (AFE), utilizando o método de análise de componentes principais (ACP) (Field, 2009). Os pressupostos para a realização desta análise foram aferidos utilizando o

KMO (*Kaiser-Meyer-Olkin*) para análise da adequação da amostra e o teste de esfericidade de Bartlett para testar se as intercorrelações entre os itens não se configuram como uma matriz de identidade (Field, 2009). No KMO foram considerados aceitáveis os valores acima de 0,5 e bons os valores acima de 0,7 (Hutcheson e Sofroniou, 1999). No teste de esfericidade de Bartlett, as estatísticas de teste com níveis de significância  $p < 0,05$  indicam que os dados são apropriados para a realização de análise em componentes principais (Tabachnick e Fidell, 1996). O método de extração foi a ACP com rotação *varimax*. A decisão sobre o número de fatores a reter foi realizada com base no critério de Kaiser, pelo que foram considerados todos os fatores com *eigenvalues* superiores a 1. Para atribuição dos itens a cada fator, consideram-se apenas cargas fatoriais  $> 0,35$ .

A fidelidade das subescalas foi testada com recurso ao método de avaliação da consistência interna dos itens, calculando-se o Alfa de Cronbach. Em geral, consideram-se como valores aceitáveis os valores de alfa superiores a 0,70, contudo, tratando-se de um estudo exploratório, esse valor pode diminuir para um mínimo de 0,60 (Hair, Black, Babin e Anderson, 2009).

Depois de delimitada a estrutura interna das escalas, calcularam-se as pontuações totais de cada dimensão, somando os respetivos itens e dividindo este valor pelo número de itens que a integram. Começou-se por calcular a estatística descritiva para a amostra total, analisando-se ainda a distribuição destas variáveis. Seguidamente, explorou-se a relação entre as pontuações nas diferentes dimensões de valorização e utilização dos métodos de avaliação e as características demográficas e profissionais dos participantes. Relativamente às características demográficas, na variável idade, para efeito deste estudo, foram considerados dois grupos para análise: 1) professores com menos de 45 anos de idade; 2) professores com mais de 45 anos de idade. Esta alteração nas unidades de tempo relativamente ao questionário inicial permitiu obter unidades de análise mais adequadas e grupos com um número de participantes mais equilibrado (Cadime, Silva, Ribeiro e Viana, 2018). Na variável experiência profissional (variáveis profissionais) procedeu-se à agregação do tempo de serviço da seguinte forma: 1) professores com menos de 15 anos de experiência; 2) professores com mais de 15 anos de experiência. Procedeu-se igualmente à junção das categorias profissionais “professor associado com agregação” e “professor auxiliar com agregação”, e “professor associado” e “professor auxiliar”. Esta alteração nas unidades de tempo e categorias profissionais relativamente ao questionário inicial permitiu obter unidades de análise mais adequadas, tendo-se usado a data da implementação do Processo de Bolonha como marco para a definição destas unidades temporais. Para a realização destas análises, recorreu-se à análise de variância multivariada (*MANOVA*). Garantiram-se os pressupostos de independência das observações, normalidade univariada e homogeneidade das matrizes de variância-covariância (Field, 2009). Os valores do eta-quadrado parcial ( $\eta^2$ ) foram calculados como medida de tamanho do efeito, considerando as seguintes linhas orientadoras para a sua interpretação: efeito pequeno,  $\eta^2 > 0,1$ ; efeito médio,  $\eta^2 > 0,3$ ; efeito grande,  $\eta^2 > 0,5$  (Cohen, 1988).

Nos casos em que foram realizados vários testes estatísticos independentes em simultâneo procedeu-se à correção de *Bonferroni* (Field, 2009). Quando os pressupostos para a utilização de testes paramétricos não foram cumpridos, optou-se pela utilização de testes não paramétricos (Testes de *Mann-Whitney* e *Kruskal-Wallis*) (Field, 2009).

Por último, procedeu-se ao teste das correlações entre os métodos de avaliação mais valorizados e os mais utilizados, através do coeficiente de correlação de *Spearman* (Field, 2009). Este coeficiente varia entre -1 e 1, indicando a direção e força da correlação, sendo que uma maior proximidade destes extremos corresponde a uma maior associação entre as variáveis (Field, 2009), tendo-se considerado os valores de 0,1 a 0,3 e -0,1 a -0,3 fracos; os valores de 0,4 a 0,6 e -0,4 a -0,6 moderados; os valores de 0,7 a 0,9 e -0,7 a -0,9 fortes e os valores 1 e -1 perfeitos (Dancey e Reidy, 2007) na interpretação dos valores de  $r$ .

Em todas as análises foram considerados como patamares para aceitação e/ou rejeição de hipóteses nulas os valores de  $p < 0,05$  (Field, 2009).

### 3. Resultados

#### 3.1. Análise fatorial exploratória e fidelidade das subescalas

Relativamente à escala sobre valorização de métodos de avaliação, o valor de KMO permitiu comprovar a adequação da amostra (KMO=0,828) e o resultado do teste de Bartlett revelou que os dados se adequam à realização desta análise,  $\chi^2(91)=1209,61$ ,  $p < 0,001$ . De igual modo, na escala de métodos de avaliação mais utilizados, o valor de KMO foi também elevado, sugerindo a adequação da amostra (KMO=0,762), e o resultado do teste de Bartlett revelou que as intercorrelações entre os itens são suficientemente elevadas para a realização desta análise,  $\chi^2(91)=836,67$ ,  $p < 0,001$ .

O quadro 2 reporta os resultados da AFE para a valorização e utilização de métodos de avaliação pelos participantes no estudo.

Uma primeira extração com a escala sobre a valorização de métodos de avaliação revelou a presença de três fatores com *eigenvalues* > 1, explicando no total 64,24% da variância. Os itens que se agruparam no mesmo fator sugerem que o fator 1 representa os Métodos Coletivos e Individuais; que o fator 2 representa os Portefólios e Reflexões; e, que o fator 3 representa os Testes e Exames. Os valores do Alfa de Cronbach foram elevados nos fatores 1 ( $\alpha=0,880$ ) e 2 ( $\alpha=0,850$ ), mas pobres para o fator 3 ( $\alpha=0,448$ ), tendo-se optado pela exclusão dos itens que compunham este fator (1-Testes/exames escritos e 2-Testes/exames orais). Contudo, dada a relevância do item 1, Testes/Exames escritos -, ao nível das respostas dos participantes do estudo e também ao nível dos estudos nacionais e internacionais (e.g. Barreira et al., 2015; Flores et al., 2014; Myers e Myers, 2015; Pereira, Flores e Barros, 2017; Pereira e Flores, 2016), optou-se pelo seu tratamento enquanto variável observável.

Após a remoção dos dois itens referidos, procedeu-se a uma nova ACP, utilizando os restantes itens. Os resultados desta análise são apresentados no quadro 2. Obtiveram-se três fatores com *eigenvalues* superiores a 1, explicando 69,17% da variância: F<sub>1</sub>) Métodos Coletivos; F<sub>2</sub>) Portefólios e Reflexões; e F<sub>3</sub>) Métodos Individuais. Todos os itens apresentaram saturações fatoriais elevadas no respetivo fator (ver quadro 2) e valores de Alfa de Cronbach > 0,70. Os itens 9, 11 e 12 saturaram simultaneamente em dois fatores tendo-se optado pela sua permanência no fator em que apresentaram maior carga fatorial (Tabachnik, e Fidell, 1996), dado que o seu conteúdo estava de acordo com o fator.

Na frequência da utilização de métodos de avaliação, uma primeira extração revelou a presença de quatro fatores. Os quatro fatores extraídos na análise foram: 1) Métodos Coletivos; 2) Portefólios e Reflexões; 3) Métodos Individuais; e 4) Testes e Exames. A

estatística de fidelidade, calculada através do Alfa de Cronbach revelou a existência de valores aceitáveis para os fatores 1 ( $\alpha=0,822$ ), 2 ( $\alpha=0,772$ ) e 3 ( $\alpha=0,693$ ). Contudo, os valores do Alfa revelaram-se inaceitáveis para o fator 4 ( $\alpha=0,342$ ), tendo-se optado pela sua exclusão. À semelhança da escala anterior, optou-se igualmente pelo tratamento do item 1 - “Testes/Exames escritos” -, em análises subsequentes, enquanto variável observável devido à sua relevância. Procedeu-se a uma nova extração de fatores, após a remoção dos itens 1 e 2. Os resultados da segunda análise são apresentados no quadro 2, através de uma estrutura de 3 fatores que explica 61,15% da variância total. Todos os itens revelaram saturações muito elevadas nos respetivos fatores. Os itens 5, 8, 11 e 12 apresentaram saturações em mais que um fator, tendo-se optado pela sua inclusão no fator com maior carga fatorial (Tabachnik e Fidell, 1996), dado que o seu conteúdo era congruente com o fator.

Quadro 2. Resultados da AFE para a valorização de métodos de avaliação e a frequência da utilização de métodos de avaliação

	FATORES – VALORIZAÇÃO DE MÉTODOS DE AVALIAÇÃO			FATORES – FREQUÊNCIA DA UTILIZAÇÃO DE MÉTODOS DE AVALIAÇÃO		
	1 MÉTODOS COLETIVOS E REFLEXÕES INDIVIDUAIS	2 PORTEFÓLIOS	3 MÉTODOS INDIVIDUAIS	1 MÉTODOS COLETIVOS E REFLEXÕES INDIVIDUAIS	2 PORTEFÓLIOS	3 MÉTODOS INDIVIDUAIS
6 - Trabalhos práticos ou experimentais em grupo	,808			,848		
8 - Projeto realizado em grupo	,713			,622		
10 - Relatórios em grupo	,789			,790		
14 - Apresentações orais em grupo	,730			,733		
3 - Portefólios coletivos		,770			,800	
4 - Portefólios individuais		,840			,811	
11 - Reflexões escritas individuais		,726			,558	
12 - Reflexões escritas em grupo		,666			,635	
5 - Trabalhos práticos ou experimentais individuais			,740			,565
7 - Projeto realizado individualmente			,718			,745
9 - Relatórios individuais			,660			,660
13 - Apresentações orais individuais			,756			,733
Eigenvalues	6.064	1.170	1.067	4.549	1.472	1.317
% Variância	50,53%	9,75%	8,90%	37,91%	12,26%	10,97%
Alfa de Cronbach	,848	,850	,807	,822	,772	,693

Nota. Os itens 1) Testes/exames escritos e 2) Testes/exames orais foram removidos nas duas escalas.

Fonte: Elaboração própria.

### 3.2. Diferenças nos métodos de avaliação mais valorizados e utilizados em função de variáveis demográficas e profissionais

O quadro 3 apresenta os resultados da estatística descritiva relativos à valorização e utilização de métodos de avaliação. Os testes de normalidade revelaram-se estatisticamente significativos, sugerindo a não normalidade das distribuições. Por conseguinte, foi dada prioridade à análise dos valores de assimetria e curtose (ver quadro 3). Os valores de assimetria e curtose são muito próximos de zero para todos os fatores de

ambas as escalas, sugerindo a não existência de violações substanciais da normalidade, o que permitiu avançar com a análise MANOVA. Contudo, verificou-se que os valores de assimetria e curtose da variável observável “testes e exames escritos” eram elevados, o que impossibilitou a inclusão destas pontuações nas análises de variância multivariada, tendo-se optado pela realização de testes não paramétricos para a sua análise.

Os “Testes/Exames Escritos” surgem como os métodos mais valorizados pelos professores, enquanto os “Portefólios e Reflexões” são os menos valorizados (ver quadro 3). Os resultados das estatísticas descritivas revelam também uma tendência positiva na valorização quer dos métodos individuais quer dos métodos coletivos.

Quadro 3. Estatísticas descritivas das escalas métodos de avaliação mais valorizados e utilizados

	N	%	MÉDIA	DESVIO PADRÃO	ASSIMETRIA	CURTOSE	TESTE KS	
							KS	P
<b>Valorização de métodos de avaliação</b>								
Fator 1 - Métodos coletivos			2,60	0,881	-0,333	-0,875	,167	<,001
Fator 2 - Portefólios e Reflexões	163	88,1	2,05	0,901	0,393	-1,067	,139	<,001
Fator 3 - Métodos Individuais			2,67	0,868	-0,399	-0,763	,152	<,001
Testes/Exames Escritos			3,24	0,784	-0,917	0,565	,251	<,001
<b>Frequência da utilização de métodos de avaliação</b>								
Fator 1 - Métodos coletivos			2,58	0,857	-0,217	-0,777	0,108	<,001
Fator 2 - Portefólios e Reflexões	171	92,4	1,82	0,736	0,797	0,156	0,132	<,001
Fator 3 - Métodos Individuais			2,44	0,714	-0,071	-0,444	0,108	<,001
Testes/Exames Escritos			3,37	0,811	-1,106	0,425	0,332	<,001

**Nota.** KS = Kolmogorov-Smirnov.

Fonte: Elaboração própria.

### 3.2.1. Influência das variáveis demográficas na valorização e utilização de métodos de avaliação

Os testes multivariados permitiram verificar que não existe uma influência da idade na valorização (WILK'S  $\Delta=,969$ ,  $F(3, 159)=1.683$ ,  $p=,173$ ;  $\eta^2=,031$ ) e na utilização dos diferentes métodos de avaliação (WILK'S  $\Delta=,960$ ,  $F(3, 167)=2.339$ ,  $p=,075$ ,  $\eta^2=,040$ ), assim como do sexo ao nível da valorização (WILK'S  $\Delta=,965$ ,  $F(3, 136)=1.652$ ,  $p=,180$ ;  $\eta^2=,035$ ) e da utilização dos diferentes métodos de avaliação (WILK'S  $\Delta=,980$ ,  $F(3, 143)=,993$ ,  $p=,398$ ,  $\eta^2=,020$ ). Também não foram encontradas diferenças estatisticamente significativas entre homens e mulheres ( $U=3217$ ,  $p=,994$ ), nem entre grupos etários ( $U=4026$ ,  $p=,635$ ) na valorização de testes e exames escritos. De igual modo, não foram identificadas diferenças estatisticamente significativas entre homens e mulheres, ( $U=2861,50$ ,  $p=,173$ ) e entre grupos etários ( $U=4047$ ,  $p=,588$ ) na frequência da utilização de testes e exames escritos.

### 3.2.2. Influência das variáveis profissionais na valorização e utilização de métodos de avaliação

Os resultados da MANOVA revelaram não existir efeitos estatisticamente significativos da categoria profissional (WILK'S  $\Delta=,963$ ,  $F(9.382)=,658$ ,  $p=,747$ ,  $\eta^2=,012$ ), dos anos de experiência (WILK'S  $\Delta=,977$ ,  $F(3, 159)=1.220$ ,  $p=,304$ ,  $\eta^2=,023$ ), e da frequência de formação pedagógica (WILK'S  $\Delta=,967$ ,  $F(3, 156)=1.756$ ,  $p=,158$ ,  $\eta^2=,033$ ) na valorização dos diferentes métodos de avaliação. De igual modo, os resultados da análise

multivariada indicaram não existir influência da categoria profissional (WILK'S  $\Delta=,926$ ,  $F(9, 402)=1,430$ ,  $p=,173$ ,  $\eta^2=,025$ ), anos de experiência (WILK'S  $\Delta=,968$ ,  $F(3, 167)=1,854$ ,  $p=,139$ ,  $\eta^2=,032$ ), e da frequência de formação pedagógica (WILK'S  $\Delta=,954$ ,  $F(3, 163)=2,632$ ,  $p=,052$ ,  $\eta^2=,046$ ), na utilização dos diferentes métodos de avaliação.

No que concerne à influência dos ciclos de estudos na valorização de métodos de avaliação, os testes multivariados revelaram não existir diferenças estatisticamente significativas ao nível dos cursos de Licenciatura (WILK'S  $\Delta=,967$ ,  $F(3, 159)=1,794$ ,  $p=,150$ ,  $\eta^2=,033$ ) e de Doutoramento, (WILK'S  $\Delta=,971$ ,  $F(3, 155)=1,569$ ,  $p=,199$ ,  $\eta^2=,029$ ). Não obstante, os resultados da MANOVA permitiram verificar a existência de diferenças estatisticamente significativas, embora com tamanho do efeito negligível, entre professores que lecionam nos cursos de Mestrado Integrado e os que não lecionam, WILK'S  $\Delta=,913$ ,  $F(3, 151)=4,786$ ,  $p=,003$ ,  $\eta^2=,087$ . Os testes univariados indicam que os professores que lecionam neste ciclo de estudos, em média, valorizam menos os portefólios e reflexões do que os professores que lecionam apenas noutros ciclos (ver quadro 4). Os testes multivariados sugeriram também diferenças significativas, embora com tamanho do efeito negligível, na valorização dos diferentes métodos de avaliação entre os professores que lecionam em cursos de Mestrado e aqueles que não lecionam, WILK'S  $\Delta=,922$ ,  $F(3, 159)=4,508$ ,  $p=,005$ ,  $\eta^2=,078$ . Os resultados dos testes univariados permitiram concluir que os professores que lecionam neste ciclo de estudos valorizam mais os métodos coletivos, individuais e portefólios e reflexões do que aqueles que não lecionam neste ciclo de estudos (ver quadro 4).

Relativamente à utilização dos diferentes métodos de avaliação de acordo com os diferentes ciclos de estudos, os testes multivariados revelaram não existir diferenças estatisticamente significativas na utilização dos diferentes métodos de avaliação pelos professores que lecionam nos cursos de Licenciatura e aqueles que não lecionam, WILK'S  $\Delta=,959$ ,  $F(3, 167)=2,366$ ,  $p=,073$ ,  $\eta^2=,041$ . Não obstante, os resultados da MANOVA permitiram verificar a existência de diferenças estatisticamente significativas, com tamanho do efeito pequeno, entre professores que lecionam nos cursos de Mestrado Integrado e os que não lecionam, WILK'S  $\Delta=,852$ ,  $F(3, 158)=9,174$ ,  $p<,001$ ,  $\eta^2=,148$ . Os professores que lecionam neste ciclo de estudos utilizam, em média, com maior frequência os métodos coletivos e com menor frequência os portefólios e reflexões (ver quadro 4). Os testes multivariados sugeriram também diferenças significativas na frequência da utilização dos diferentes métodos de avaliação entre os professores que lecionam no Mestrado e aqueles que não lecionam, WILK'S  $\Delta=,898$ ,  $F(3,167)=6,293$ ,  $p<,001$ ,  $\eta^2=,102$ , bem como entre aqueles que lecionam nos cursos de Doutoramento e os que não lecionam, WILK'S  $\Delta=,949$ ,  $F(3, 163)=2,910$ ,  $p=,036$ ,  $\eta^2=,051$ . Em ambos os casos, os resultados dos testes univariados permitiram concluir que os professores que lecionam nestes ciclos de estudos utilizam com maior frequência os métodos de avaliação coletivos, individuais e portefólios e reflexões do que aqueles que não lecionam nesses mesmos ciclos de estudo (ver quadro 4).

Quadro 4. Resultados dos testes univariados para a análise do efeito do Ciclo de Estudos na valorização e frequência da utilização de métodos de avaliação

	LICENCIATURA					MESTRADO INTEGRADO					MESTRADO					DOUTORAMENTO				
	Sim M <sub>e</sub> (DP)	Não M <sub>e</sub> (DP)	F (gl)	P	ηp <sup>2</sup>	Sim M <sub>e</sub> (DP)	Não M <sub>e</sub> (DP)	F (gl)	P	ηp <sup>2</sup>	Sim M <sub>e</sub> (DP)	Não M <sub>e</sub> (DP)	F (gl)	P	ηp <sup>2</sup>	Sim M <sub>e</sub> (DP)	Não M <sub>e</sub> (DP)	F (gl)	P	ηp <sup>2</sup>
<b>Valorização de métodos de avaliação</b>																				
<b>Métodos Coletivos</b>	2,66 (,87)	2,32 (,92)	3,38 (1,16)	,07	,02	2,72 (,77)	2,56 (,96)	1,29 (1,15)	,26	,008	2,72 (,84)	2,18 (,90)	10,63 (1,16)	,001	,06	2,74 (,83)	2,46 (,92)	3,24 (1,16)	,04	,03
<b>Portefólios e reflexões</b>	2,11 (,90)	1,71 (,85)	4,69 (1,16)	,03	,03	1,88 (,84)	2,20 (,94)	4,59 (1,15)	,03	,03	2,13 (,89)	1,76 (,89)	4,63 (1,16)	,03	,03	2,16 (,87)	1,94 (,94)	1,80 (1,16)	,13	,01
<b>Métodos Individuais</b>	2,73 (,84)	2,38 (,98)	3,66 (1,16)	,06	,03	2,69 (,80)	2,72 (,91)	0,04 (1,15)	,84	,00	2,78 (,82)	2,25 (,93)	11,37 (1,16)	,001	,06	2,80 (,75)	2,55 (,98)	2,48 (1,16)	,07	,02
<b>Frequência da utilização de métodos de avaliação</b>																				
<b>Métodos Coletivos</b>	2,60 (,84)	2,49 (,94)	0,39 (1,17)	,53	,002	2,77 (,75)	2,42 (,91)	6,89 (1,16)	,009	,04	2,67 (,85)	2,27 (,82)	6,95 (1,17)	,009	,04	2,73 (,84)	2,40 (,87)	6,23 (1,16)	,01	,04
<b>Portefólios e reflexões</b>	1,88 (,75)	1,54 (,61)	5,18 (1,17)	,02	,03	1,66 (,68)	1,90 (,71)	4,67 (1,16)	,03	,03	1,93 (,76)	1,46 (,52)	13,56 (1,17)	<,001	,07	1,94 (,73)	1,68 (,73)	5,05 (1,16)	,03	,03
<b>Métodos Individuais</b>	2,48 (,69)	2,21 (,77)	3,69 (1,17)	,06	,02	2,38 (,70)	2,49 (,72)	0,99 (1,16)	,32	,006	2,54 (,70)	2,09 (,68)	12,77 (1,17)	<,001	,07	2,56 (,65)	2,31 (,77)	5,28 (1,16)	,02	,03

Nota. M<sub>e</sub> (média); DP (Desvio Padrão); gl (graus de liberdade)

Fonte: Elaboração própria.

Relativamente à área de conhecimento em que os professores lecionam, os testes multivariados revelaram a existência de diferenças estatisticamente significativas, embora com tamanho do efeito negligível, na valorização dos diferentes métodos de avaliação (WILK'S  $\Delta=,744$ ,  $F(12, 413)=4,063$ ,  $p<,001$ ,  $\eta^2=,094$ ) e de diferenças estatisticamente significativas, com tamanho de efeito pequeno na utilização dos diferentes métodos de avaliação WILK'S  $\Delta=,694$ ,  $F(12, 434)=5,360$ ,  $p<,001$ ,  $\eta^2=,115$ .

Os resultados dos testes univariados (quadro 5) e as comparações *pairwise* correspondentes permitiram identificar uma maior valorização de métodos coletivos nos professores que lecionam nas áreas das Ciências da Engenharia e Tecnologia em relação aos professores das Ciências Exatas ( $p=,001$ ) e Ciências Médicas e da Saúde ( $p=,047$ ); e nos professores das Ciências Sociais em relação aos professores das Ciências Exatas ( $p=,008$ ). Permitiram igualmente identificar uma maior valorização de portefólios e reflexões pelos docentes das Ciências Sociais em relação aos docentes das Ciências Exatas ( $p=,007$ ) e das Ciências da Engenharia e Tecnologia ( $p=,049$ ). As restantes comparações não foram estatisticamente significativas.

No que diz respeito à utilização dos métodos de avaliação, os testes univariados identificaram a existência de diferenças estatisticamente significativas entre professores de diferentes áreas de conhecimento, porém com tamanho do efeito negligível, na utilização de portefólios e reflexões e de métodos coletivos (ver quadro 5), favoráveis, de acordo com as comparações *pairwise*, à utilização de métodos coletivos pelos docentes das Ciências da Engenharia e Tecnologia em relação aos docentes das Ciências Exatas ( $p<,001$ ) e Humanidades ( $p=,004$ ); e dos docentes das Ciências Sociais em relação aos docentes das Ciências Exatas ( $p=,001$ ). Relativamente à utilização de portefólios e reflexões, as comparações *pairwise* revelaram uma maior utilização destes métodos pelos docentes das Ciências Sociais em relação aos docentes das Ciências Exatas ( $p=,015$ ), das Ciências Médicas e da Saúde ( $p=,009$ ) e das Ciências da Engenharia e Tecnologia ( $p=,005$ ).

Não foram encontradas diferenças estatisticamente significativas entre a experiência profissional dos participantes ( $U=3330$ ,  $p=,550$ ); os ciclos de estudo, Licenciatura ( $U=2090,50$ ,  $p=,369$ ), Mestrado Integrado ( $U=3647$ ,  $p=,799$ ), Mestrado ( $U=2689$ ,  $p=,378$ ) e Doutoramento ( $U=3570$ ,  $p=,184$ ); e a categoria profissional ( $\chi^2=0,990(2)$ ,  $p=609$ ), na valorização de testes e exames escritos. Contudo, foram encontradas diferenças estatisticamente significativas ao nível da área científica e da frequência de formação pedagógica (ver quadro 6). A valorização dos testes e exames escritos é significativamente maior nos professores que não possuem formação pedagógica. Relativamente às diferenças identificadas ao nível da área científica, foram aplicados testes *Mann-Whitney* subsequentes para investigar as diferenças entre grupos (*pairwise*). Foi aplicada uma correção de *Bonferroni* e adotado o nível de significância de  $p<,005$  (valor resultante da divisão do valor de significância,05 pelo número de testes realizados (10)). Os professores das Ciências Exatas reportaram uma maior valorização de testes e exames escritos do que os professores das Ciências Sociais ( $U=307,50$ ,  $p=,002$ ) e os professores das Humanidades reportaram uma maior valorização em relação aos professores das Ciências Sociais ( $U=391,50$ ,  $p<,001$ ). As restantes comparações entre grupos não foram significativas.

Quadro 5. Resultados dos testes univariados para a análise do efeito da área de conhecimento na valorização e frequência da utilização de métodos de avaliação

	ÁREA DE CONHECIMENTO						P	$\eta^2$
	CE M <sub>e</sub> (DP)	CET M <sub>e</sub> (DP)	CMS M <sub>e</sub> (DP)	CS M <sub>e</sub> (DP)	H M <sub>e</sub> (DP)	F (gl)		
<b>Valorização de métodos de avaliação</b>								
Métodos	1,89	2,90	2,26	2,72	2,28	5,919	<,001	,130
Coletivos	(,897)	(,644)	(,926)	(,887)	(,895)	(4,158)		
Portefólios e reflexões	1,46	1,87	1,83	2,35	2,11	4,361	,0,02	,099
	(,611)	(,828)	(,946)	(,948)	(,698)	(4,158)		
Métodos Individuais	2,23	2,76	2,51	2,74	2,69	1,280	,280	,031
	(,968)	(,824)	(,985)	(,864)	(,735)	(4,158)		
<b>Frequência da utilização de métodos de avaliação</b>								
Métodos	1,79	2,90	2,41	2,69	2,12	7,582	<,001	,154
Coletivos	(,587)	(,729)	(,834)	(,862)	(,805)	(4,166)		
Portefólios e reflexões	1,46	1,66	1,51	2,12	1,68	6,007	<,001	,126
	(,489)	(,689)	(,580)	(,817)	(,352)	(4,166)		
Métodos Individuais	2,20	2,35	2,24	2,56	2,57	1,689	,155	,039
	(,735)	(,676)	(,729)	(,722)	(,701)	(4,166)		

Nota. M<sub>e</sub> (média); DP (Desvio Padrão); gl (graus de liberdade), CE (Ciências Exatas), CET (Ciências da Engenharia e da Tecnologia); CMS (Ciências Médicas e da Saúde), CS (Ciências Sociais), H (Humanidades).

Fonte: Elaboração própria.

Ao nível da frequência da utilização de testes e exames escritos não foram igualmente encontradas diferenças estatisticamente significativas entre a experiência profissional dos participantes ( $U=3331$ ,  $p=,488$ ); os ciclos de estudo de Licenciatura ( $U=2215$ ,  $p=,648$ ), Mestrado Integrado ( $U=3501$ ,  $p=,385$ ), Mestrado ( $U=2810$ ,  $p=,481$ ) e Doutorado ( $U=3935$ ,  $p=,737$ ); e a categoria profissional ( $\chi^2= 2.525(2)$ ,  $p=,283$ ). Porém, foram encontradas diferenças estatisticamente significativas ao nível da área científica e da frequência de ações de formação na utilização de testes e exames escritos (ver quadro 6). A valorização dos testes e exames escritos é significativamente maior nos professores que não possuem formação pedagógica. Relativamente às diferenças identificadas ao nível da área de conhecimento, foram aplicados testes *Mann-Whitney* subsequentes para investigar as diferenças entre grupos (*pairwise*). Foi aplicada uma correção de *Bonferroni* e adotado o nível de significância de  $p<.005$  (valor resultante da divisão do valor de significância  $p<.05$  pelo número de testes realizados (10)). Os professores das Ciências Exatas reportaram uma maior frequência da utilização de testes e exames escritos em relação aos professores das Ciências da Engenharia e Tecnologia ( $U=280.50$ ,  $p=,023$ ), aos professores das Ciências Médicas e da Saúde ( $U=97.50$ ,  $p=,029$ ) e aos professores das Ciências Sociais ( $U=282.50$ ,  $p<,001$ ); e os professores das Ciências da Engenharia e Tecnologia reportaram uma maior frequência da utilização de testes e exames escritos em relação aos professores das Ciências Sociais ( $U=1298$ ,  $p=,001$ ). As restantes comparações entre grupos não foram significativas.

Tabela 6. Resultados dos testes não paramétricos para a análise do efeito da frequência de ações de formação pedagógica, área de conhecimento e indicação de práticas de avaliação na valorização e utilização de testes e exames escritos

	FREQUÊNCIA DE AÇÕES DE FORMAÇÃO PEDAGÓGICA				ÁREA DE CONHECIMENTO						INDICAÇÃO DE ALTERAÇÃO DE PRÁTICAS DE AVALIAÇÃO AO LONGO DA CARREIRA					
	Sim PM	Não PM	U	P	CE PM	CET PM	CMS PM	CS PM	H PM	$\chi^2$ (gl)	P	Sim	Não	Talvez	$\chi^2$ (gl)	P
<b>Valorização de testes e exames escritos</b>	83,98	102,60	2923	,012	120,60	96,28	98,95	74,42	123,29	24,64 (4)	<,001	88,58	119,44	97,44	6,68 (2)	,035
<b>Frequência da utilização de testes e exames escritos</b>	82,72	104,95	2775	,002	129,63	105,64	92,26	76,19	97,36	22,35 (4)	<,001	96,34	90,82	83,33	2,53 (2)	,283

Nota. PM (ponto médio); U (U de Mann-Whitney);  $\chi^2$  (Qui-quadrado); gl (graus de liberdade), CE (Ciências Exatas), CET (Ciências da Engenharia e da Tecnologia); CMS (Ciências Médicas e da Saúde), CS (Ciências Sociais), H (Humanidades)

Fonte: Elaboração própria.

### 3.2.3. Influência das variáveis relacionadas com as práticas de avaliação e com o Processo de Bolonha na valorização e utilização de métodos de avaliação

Um aspeto importante no contexto da investigação era compreender se os professores que afirmam ter alterado a forma como avaliam os seus alunos valorizam e utilizam métodos de avaliação diferentes dos professores que responderam negativamente. Os resultados da MANOVA revelaram não existir uma influência da mudança das práticas de avaliação na valorização (WILK'S  $\Delta=,931$ ,  $F(6, 314)= 1.905$ ,  $p=.80$ ,  $\eta^2=,035$ ) e na frequência da utilização (WILK'S  $\Delta=,951$ ,  $F(6, 330)=1.393$ ,  $p=,217$ ,  $\eta^2=,025$ ) dos diferentes métodos de avaliação. De igual modo, os resultados dos testes multivariados não permitiram identificar uma influência do reconhecimento do papel do Processo de Bolonha na mudança das práticas de avaliação na valorização (WILK'S  $\Delta=,985$ ,  $F(6, 310)=,402$ ,  $p=.878$ ,  $\eta^2=,008$ ) e na frequência da utilização (WILK'S  $\Delta=,949$ ,  $F(6, 326)=1,430$ ,  $p=.202$ ,  $\eta^2=,026$ ) dos diferentes métodos de avaliação.

Não foram encontradas diferenças estatisticamente significativas entre o reconhecimento do papel do Processo Bolonha na mudança das práticas de avaliação no Ensino Superior ( $\chi^2=0,101(2)$ ,  $p=,951$ ) na valorização de testes e exames escritos. Contudo, foram encontradas diferenças estatisticamente significativas ao nível da mudança de práticas de avaliação na valorização de testes e exames escritos (ver quadro 6). Foram aplicados testes *Mann-Whitney* para investigar as diferenças entre grupos. Foi aplicada uma correção de *Bonferroni* e adotado o nível de significância de  $p<,016$  (valor resultante da divisão do valor de significância  $p<.05$  pelo número de testes realizados (3)). Os professores que não alteraram as suas práticas de avaliação no Ensino Superior reportaram uma maior valorização de testes e exames escritos em relação aos professores que responderam afirmativamente ( $U=935.50$ ,  $p=,01$ ). As restantes comparações entre grupos não foram significativas.

Relativamente à frequência da utilização de testes e exames escritos não foram encontradas diferenças estatisticamente significativas ao nível do da influência do Processo Bolonha na mudança das práticas de avaliação no Ensino Superior ( $\chi^2=0.104(2)$ ,  $p=,949$ ), assim como da mudança de práticas de avaliação (ver quadro 6).

### 3.3. Correlação entre os métodos de avaliação mais valorizados e os métodos de avaliação mais utilizados

O quadro 7 apresenta a matriz de correlações entre a valorização e a frequência da utilização de métodos de avaliação. Foi identificada uma correlação significativa forte entre a valorização e frequência da utilização dos diferentes métodos de avaliação: uma maior valorização de métodos coletivos, portefólios e reflexões, e métodos individuais está associada a uma maior frequência da utilização dos mesmos. Foi também encontrada uma relação positiva e moderada entre a valorização e a frequência da utilização de testes e exames escritos.

Não obstante, a análise dos resultados do coeficiente de *Spearman* permitem identificar que uma maior valorização de portefólios e reflexões está associada a menor valorização e também menor utilização de testes e exames; uma maior utilização de portefólios e reflexões está associada a menor utilização de testes e exames; uma maior utilização de métodos coletivos está associada a menor utilização de testes e exames; e uma maior

utilização de testes e exames escritos está associada a menor utilização e valorização de portfólios e reflexões; e a uma menor utilização de métodos coletivos.

Quadro 7. Matriz de correlação entre a valorização de métodos de avaliação e a frequência da utilização de métodos de avaliação (Correlação de Spearman)

	VALORIZAÇÃO DE MÉTODOS DE AVALIAÇÃO				FREQUÊNCIA DA UTILIZAÇÃO DE MÉTODOS DE AVALIAÇÃO			
	Métodos Coletivos	Portefólios e Reflexões	Métodos Individuais	Testes/exames escritos	Métodos Coletivos	Portefólios e Reflexões	Métodos Individuais	Testes/exames escritos
<b>Valorização de métodos de avaliação</b>								
Métodos Coletivos	1	,563***	,623***	-,115	,787***	,400***	,421***	-,138
Portefólios e Reflexões		1	,571***	-,172*	,482***	,822***	,489***	-,317***
Métodos Individuais			1	,054	,408***	,369***	,735***	,012
Testes/exames escritos				1	-,246**	-,295***	-,079	,584***
<b>Frequência da utilização de métodos de avaliação</b>								
Métodos Coletivos					1	,457***	,407***	-,171*
Portefólios e Reflexões						1	,409***	-,402***
Métodos Individuais							1	-,011
Testes/exames escritos								1

Nota. \* p < ,05; \*\* p < ,01; \*\*\* p < ,001.

Fonte: Elaboração própria.

#### 4. Discussão e conclusões

Este artigo visou explorar as seguintes questões de investigação: 1) Que importância atribuem aos métodos de avaliação?; 2) Qual a frequência da utilização dos diferentes métodos de avaliação?; 3) Quais as variáveis associadas à valorização e utilização de diferentes métodos de avaliação?; e 4) Qual a relação entre a valorização e a utilização de diferentes métodos de avaliação?

Para explorar estas questões, utilizou-se uma medida de valorização e outra de frequência de métodos de avaliação. Ambas as medidas demonstraram propriedades psicométricas adequadas, sendo a estrutura interna de ambas congruente com uma estrutura de três fatores: métodos coletivos, métodos individuais e portefólios e reflexões. Os itens de cada fator revelaram uma adequada consistência interna, suportando a fidelidade das pontuações obtidas. A valorização e frequência de utilização de testes/exames escritos foi analisada separadamente dada a ausência de saturação nos fatores anteriormente identificados.

Relativamente à influência das variáveis demográficas na valorização e frequência da utilização de métodos de avaliação, verificou-se não existir uma influência das variáveis demográficas (idade e sexo) ao nível da valorização e frequência da utilização dos métodos de avaliação em análise.

Ao nível das variáveis profissionais, a MANOVA permitiu identificar diferenças estatisticamente significativas em função dos ciclos de estudos na valorização e utilização de métodos coletivos, individuais, e de portfólios e reflexões; e da área de conhecimento na valorização e utilização de métodos coletivos e portfólios e reflexões. Estudos anteriores realizados com alunos do Ensino Superior português (Pereira, 2016) identificaram igualmente uma influência do curso na utilização dos diferentes métodos de avaliação. Os resultados dos testes não paramétricos permitiram identificar uma influência da formação pedagógica e da área científica ao nível da valorização e frequência da utilização de testes e exames escritos. As respostas dos participantes permitiram identificar, por um lado, uma maior valorização de testes e exames escritos pelos professores que não possuem formação pedagógica e, por outro lado, uma maior valorização dos testes e exames escritos pelos professores das Ciências Exatas e das Humanidades em relação aos professores das Ciências Sociais. Identificou-se também uma maior frequência de utilização dos testes e exames escritos pelos professores das Ciências Exatas por comparação aos professores das Ciências da Engenharia e da Tecnologia, das Ciências Médicas e da saúde e das Ciências Sociais; e dos professores das Ciências da Engenharia e da Tecnologia em comparação com os professores das Ciências Sociais.

Em relação à influência das variáveis relacionadas com as práticas de avaliação e com o Processo de Bolonha na valorização e frequência da utilização de métodos de avaliação, não se verificou uma influência destas variáveis na valorização e utilização de métodos individuais, coletivos e de portfólios e reflexões. Porém, verificou-se a influência da mudança de práticas de avaliação ao nível da valorização de testes e exames escritos, sugerindo uma maior valorização de testes e exames pelos professores que afirmam não ter alterado as suas práticas de avaliação ao longo da sua carreira.

Na análise das correlações entre a valorização e a frequência da utilização de métodos de avaliação destaca-se a correlação positiva forte entre a valorização e utilização dos métodos de avaliação. Contudo, essa correlação é moderada ao nível da valorização e utilização de testes e exames escritos.

Estes resultados apontam para a valorização e frequência da utilização de um leque diversificado de métodos de avaliação pelos participantes neste estudo. Porém, os resultados das estatísticas descritivas apontam para uma maior valorização e frequência de utilização de testes e exames escritos e para uma menor valorização e frequência de utilização de portfólios e reflexões. Estes resultados corroboram outros estudos realizados no contexto português (Barreira et al., 2015; Pereira, 2016) que destacam o carácter sumativo da avaliação e a prevalência do uso de testes e exames escritos, articulada com a utilização de outros métodos de avaliação.

Estudos anteriores no contexto português apontam para a existência de contradições entre as práticas de avaliação utilizadas e as concepções de avaliação dos docentes (Gonçalves, 2016; Pereira e Flores, 2016) justificadas pelo volume de trabalho, escassez de recursos humanos e físicos e a imposição institucional na utilização de avaliação sumativa, que perpetuam o uso de determinados métodos de práticas (Pereira e Flores, 2016). Os resultados deste estudo indicam alguma coerência entre a valorização dos diferentes métodos de avaliação e a frequência da sua utilização, todavia, com menor expressividade na valorização e frequência da utilização de testes e exames escritos, o que poderá ser um resultado da imposição e prevalência da avaliação sumativa (Pereira e Flores, 2016).

Estes resultados realçam a complexidade dos cenários formativos universitários (Zabalza, 2004), da avaliação (Brown e Knight, 1994) e da profissão docente no sentido de encarar a docência como um processo de criação e desenvolvimento de conhecimento através do estudo e exploração das suas diferentes dimensões, nomeadamente, o desenvolvimento profissional docente com particular ênfase em práticas de avaliação inovadoras (Fernandes, 2015).

Os resultados deste estudo contribuem para compreender o processo de avaliação na ótica de professores universitários portugueses, sugerindo-se a aplicação futura do instrumento a um número mais expressivo de docentes, de diferentes contextos universitários, para melhor conhecer as suas perceções acerca dos métodos e práticas de avaliação.

## Agradecimentos

Este estudo é financiado por Fundos Nacionais através da FCT (Fundação para a Ciência e a Tecnologia) e cofinanciado pelo Fundo Europeu de Desenvolvimento Regional (FEDER) através do COMPETE 2020 – Programa Operacional Competitividade e Internacionalização (POCI) com a referência POCI-01-0145-FEDER-007562 no âmbito do projeto “Assessment in Higher Education: the potential of alternative methods”, com a referência PTDC/MHCCED/2703/2014, e do projeto de doutoramento em Ciências da Educação - Especialidade em Desenvolvimento Curricular intitulado “Conceções e práticas de avaliação no Ensino Superior após a implementação do Processo de Bolonha: um estudo com professores universitários”, com a referência SFRH/BD/103291/2014.

O projeto é desenvolvido no Centro de Investigação em Estudos da Criança, no âmbito do Projeto Estratégico UID/ CED/00317/2013, por Fundos Nacionais através da FCT (Fundação para a Ciência e Tecnologia) e co-financiado pelos Fundos Europeus de Desenvolvimento Regional (FEDER), através do Programa Operacional de Competitividade e Internacionalização (POCI) com a referência POCI-01-0145-FEDER-007562.



## Referências

- Barreira, C., Bidarra, M. G., Vaz-Rebelo, P. Monteiro, F. e Alferes, V. (2015). Perceções dos professores e estudantes de quatro universidades portuguesas acerca do ensino e avaliação das aprendizagens. En D. Fernandes, A. Borralho, C. Barreira, A. Monteiro, D. Catani, E. Cunha, e P. Alves (Orgs.), *Avaliação, ensino e aprendizagem em Portugal e no Brasil: Realidades e perspectivas* (pp. 309-326). EDUCA.
- Black, P. e Wiliam, D. (1998), Assessment and classroom learning, *Assessment in Education*, 5(1), 7-75. <https://doi.org/10.1080/0969595980050102>
- Brew, C., Riley, P. e Walta, C. (2009). Participative assessment practices: A comparison of pre-service primary teachers and teaching staff views. *Assessment and Evaluation in Higher Education*, 34(6), 641-657. <https://doi.org/10.1080/02602930802468567>
- Brown, S. e Knight, P. (1994). *Assessing learners in higher education*. Kogan Page.

- Burkšaitienė, N. e Teresevičienė, M. (2008). Integrating alternative learning and assessment in a course of English for law students. *Assessment e Evaluation in Higher Education*, 33(2), 155-166. <https://doi.org/10.1080/02602930601125699>
- Cadime, I., Silva, C., Ribeiro, I. e Viana, F. L. (2018). Early lexical development: Do day care attendance and maternal education matter? *First Language*, 38(5), 503-519. <https://doi.org/10.1177/0142723718778916>
- Carless, D., Salter, M., Yang, M. e Lam, J. (2011). Developing Sustainable Feedback Practices. *Studies in Higher Education*, 36(4), 395-407. <https://doi.org/10.1080/03075071003642449>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Lawrence Earlbaum Associates.
- Coutinho, C. P. (2014). *Metodologias de investigação em ciências sociais e humanas: Teoria e prática*. Edições Almedina, S.A.
- Dancey, C. P., Reidy, J. (2007). *Statistics without maths for psychology*. Pearson Education.
- Dochy, F., Segers, M., Gijbels, D. e Struyven, K. (2007). Assessment Engineering: Breaking down barriers between teaching, learning and assessment. En D. Boud e N. Falchikov (Eds.), *Rethinking assessment in higher education: learning for the longer term* (pp. 83-100). Routledge.
- Duncan, T. e Buskirk-Cohen, A. (2011). Exploring Learner-Centered Assessment: A Cross-Disciplinary Approach. *International Journal of Teaching and Learning in Higher Education*, 23(2), 246-259.
- Fernandes, D. (2015). Práticas de ensino e de avaliação de docentes de quatro universidades portuguesas. En D. Fernandes, A. Borralho, C. Barreira, A. Monteiro, D. Catani, E. Cunha, e P. Alves (Orgs.), *Avaliação, ensino e aprendizagem em Portugal e no Brasil: Realidades e perspectivas I* (pp. 97-135). EDUCA.
- Fernandes, S. e Flores, M. A. (2012). A docência no contexto da avaliação do desempenho no ensino superior: Reflexões no âmbito de um estudo em curso. *Revista Iberoamericana de Evaluación Educativa*, 5(2), 82-98.
- Field, A. (2009). *Discovering statistics using SPSS*. Sage Publications Ltd.
- Flores, M. A. e Veiga Simão, A. M. (2007, junho). Competências desenvolvidas no contexto do Ensino Superior: a perspectiva dos diplomados. In *V Jornadas de Redes de Investigación en Docencia Universitaria*. Alicante, Espanha.
- Flores, M. A., Veiga Simão A. M., Barros A. e Pereira, D. (2014). Perceptions of effectiveness, fairness and feedback of assessment methods: a study in higher education. *Studies in Higher Education*, 40(9), 1523-1534. <https://doi.org/10.1080/03075079.2014.881348>
- Gilles, J. L., Detroz, P. e Blais, J. G. (2010). An international online survey of the practices and perceptions of higher education professors with respect to the assessment of learning in the classroom. *Assessment e Evaluation in Higher Education* 36(6), 719-733. <https://doi.org/10.1080/02602938.2010.484880>
- Gonçalves, R. (2016). *Conceções de avaliação em contexto de ensino clínico de enfermagem: Um estudo na escola superior de enfermagem de Coimbra*. Tese de Doutoramento em Didática e Formação. Universidade de Aveiro, Portugal.
- Hadji, C. (1994). *A avaliação, regras do jogo*. Porto: Porto Editora.
- Hair, J. F., Black, W. C., Babin, B. J. e Anderson, R. E. (2009). *Multivariate data analysis*. Prentice Hall.
- Hutcheson, G. D. e Sofroniou, N. (1999). *The multivariate social scientist: Introductory statistics*. Sage Publications.

- Maclellan, E. (2004). How convincing is alternative assessment for use in higher education?. *Assessment e Evaluation in Higher Education*, 29(3), 311-321. <https://doi.org/10.1080/0260293042000188267>
- McDowell, L., Wakelin, D., Montgomery, C. e King, S. (2011). Does assessment for learning make a difference? The development of a questionnaire to explore the student response. *Assessment e Evaluation in Higher Education*, 36(7), 749-765. <https://doi.org/10.1080/02602938.2010.488792>
- Myers, C. B. e Myers, S. M. (2015). The use of learner-centered assessment practices in the United States: the influence of individual and institutional contexts. *Studies in Higher Education*, 40(10), 1904-1918. <https://doi.org/10.1080/03075079.2014.914164>
- Pereira, D. R. (2011). *A avaliação das aprendizagens no ensino superior na perspetiva dos estudantes. Um estudo exploratório*. Dissertação de Mestrado em Ciências da Educação. Universidade do Minho, Portugal.
- Pereira, D. R. (2016). *Assessment in higher education and quality of learning: Perceptions, practices and implications*. Tese de Doutoramento em Ciências da Educação. Universidade do Minho, Portugal.
- Pereira, D. R. e Flores, M. A. (2013). Avaliação e feedback no ensino superior: um estudio na Universidade do Minho. *Revista Iberoamericana de Educación Superior*, 4(10), 40-54.
- Pereira, D. R. e Flores, M. A. (2016). Conceptions and practices of assessment in Higher Education: A study of Portuguese university teachers. *Revista Iberoamericana de Evaluación Educativa*, 9(1), 9-29. <https://doi.org/10.15366/riee2016.9.1.001>
- Pereira, D., Flores, M. A. e Barros, A. (2017). Perceptions of Portuguese undergraduate students about assessment: a study in five public universities. *Educational Studies*, 43(4), 442-463. <https://doi.org/10.1080/03055698.2017.1293505>
- Perrenoud, P. (1999). *Avaliação: da excelência à regulação das aprendizagens: entre duas lógicas*. Porto Alegre: Artmed.
- Phelps, R. P. (2017). The “teaching to the test” family of fallacies. *Revista Iberoamericana de Evaluación Educativa*, 10(1), 33-49. <https://doi.org/10.15366/riee2017.10.1.002>
- Segers, M., Gijbels, D. e Thurlings, M. (2008). The relationship between students' perceptions of portfolio assessment practice and their approaches to learning. *Educational Studies*, 34(1), 35-44. <https://doi.org/10.1080/03055690701785269>
- Struyven, k., Dochy, F. e Steven, J. (2005). Students' perceptions about evaluation and assessment in higher education: a review. *Assessment e Evaluation in Higher Education*, 30(4), 325-341. <https://doi.org/10.1080/02602930500099102>
- Tabachnik, B. G. e Fidell, L. S. (1996). *Using multivariate statistics*. Harper e Row.
- Watering, G., Gijbels, D., Dochy, F. e Rijt, J. (2008). Students' assessment preferences, perceptions of assessment and their relationships to study results. *Higher Education*, 56, 645-58. <https://doi.org/10.1007/s10734-008-9116-6>
- Webber, K. L. e Tschepikow, K. (2013). The role of learner-centred assessment in postsecondary organisational change. *Assessment in Education: Principles, Policy e Practice*, 20(2), 187-204. <https://doi.org/10.1080/0969594x.2012.717064>
- Zabalza, M. A. (2008). *Competencias docentes del profesorado universitario. Calidad y desarrollo profesional*. Narcea.
- Zabalza, M. A. (2004). *La enseñanza universitaria. El escenario y sus protagonistas*. Narcea.

## Breve Cv de las autoras

### **Eva Lopes Fernandes**

Eva Lopes Fernandes é aluna do doutoramento em Ciências da Educação, especialização em Desenvolvimento Curricular na Universidade do Minho, com o tema: "Conceptions and Practices of Assessment in Higher Education: A study of University Teachers" (SFRH/BD/103291/2014). É Licenciada em Educação e Mestre em Ciências da Educação pela Universidade do Minho. Tem várias publicações sobre temas relacionados com o trabalho dos professores, profissionalismo docente, liderança docente e as vozes dos alunos. Atualmente integra a equipa de investigação do projeto "Impact - Investigando os Efeitos das Lideranças Escolares nos Resultados dos Alunos", tendo integrado várias equipas de investigação de projetos nacionais e internacionais (e.g. "Assessment in Higher Education: the potential of alternative methods"; "Teachers Exercising Leadership" and "Teachers professional trajectories"). ORCID ID: <https://orcid.org/0000-0002-3838-9846>. Email: [evalopesfernandes@ie.uminho.pt](mailto:evalopesfernandes@ie.uminho.pt)

### **Maria Assunção Flores**

Professora na Universidade do Minho. Doutorou-se em Educação na Universidade de Nottingham, Reino Unido, tendo sido *visiting scholar* na Universidade de Cambridge e na Universidade de Glasgow. É membro de várias associações científicas internacionais e pertence ao corpo editorial de várias revistas internacionais. É diretora executiva da revista *Teachers and Teaching Theory and Practice* e co-diretora da *European Journal of Teacher Education*. Foi presidente da *International Study Association on Teachers and Teaching*, tendo ainda presidido à direção do *International Council on Education for Teaching*. É coordenadora e membro fundador da Rede Internacional de Investigação-Ação Colaborativa (Estreiadialogos). É ainda membro do Conselho Geral do IAVE, I.P. As suas áreas de investigação incluem formação e desenvolvimento profissional de professores, profissionalismo docente e identidade profissional, currículo, avaliação e ensino superior. Tem mais de 200 publicações nestes domínios, incluindo livros, capítulos de livros e artigos em revistas nacionais e internacionais. Coordenou vários projetos de investigação, sendo os mais recentes "Os professores e o exercício da Liderança" e "Investigando os efeitos das lideranças nos resultados escolares dos alunos", financiados pela Fundação para a Ciência e a Tecnologia. <http://orcid.org/0000-0002-4698-7483>. Email: [aflores@ie.uminho.pt](mailto:aflores@ie.uminho.pt)

### **Irene Cadime**

Licenciada e Doutorada em Psicologia, na especialidade de Psicologia da Educação, pela Universidade do Minho. Mestre em Intervenção Psicológica, Educação e Desenvolvimento Humano pela Universidade do Porto. Atualmente é investigadora e membro integrado do Centro de Investigação em Estudos da Criança do Instituto de Educação da Universidade do Minho. Desenvolve trabalhos de investigação nos domínios da psicometria, do desenvolvimento de competências linguísticas e do sucesso académico. ORCID ID: <https://orcid.org/0000-0001-8285-4824>. Email: [irenecadime@ie.uminho.pt](mailto:irenecadime@ie.uminho.pt)

### **Clara Pereira Coutinho**

Licenciada em Economia, Mestre em Educação na área de especialização de Tecnologia Educativa e Doutora em Educação na área de especialização de Tecnologia Educativa, grau que obteve no ano de 2003 na Universidade do Minho, Braga, Portugal. Atualmente é Professora Auxiliar Aposentada do Departamento de Estudos Curriculares e Tecnologia Educativa do Instituto de Educação da Universidade do Minho sendo responsável pela coordenação da linha de investigação “Recursos Educativos Digitais” no Centro de Investigação em Estudos da Criança. Tem desenvolvido atividades de pesquisa no âmbito da Formação de Professores em Tecnologias de Informação e Comunicação e ainda no domínio das Metodologias de Investigação em Educação. Mais recentemente, desenvolve investigação ao nível do *mobile learning* e da utilização de aplicativos da Web 2.0 como ferramentas de apoio ao ensino e à aprendizagem, tendo publicados dezenas de artigos em revistas internacionais de referência bem como em atas de reuniões científicas nacionais e internacionais. Participa em diversos projetos financiados por agências nacionais e internacionais tendo recebido diversos prémios, como é o caso do projeto *t-words*, um manipulativo digital para crianças do pré-escolar que recebeu o World Technology Award em 2013. Publicou dois livros um dos quais no domínio das Metodologias de Investigação em Ciências Sociais e Humanas que teve a sua 2ª edição publicada em 2013. Informação adicional pode ser encontrada na sua página pessoal disponível em <http://www.degois.pt/visualizador/curriculum.jsp?key=1426606078182665>. ORCID ID: <http://orcid.org/0000-0002-2309-4084>. Email: [ccoutinho@ie.uminho.pt](mailto:ccoutinho@ie.uminho.pt)

# Evaluación basada en los Resultados de Aprendizaje: Una Experiencia en la Universidad

## Learning Outcomes based Assessment: An Experience at University

Eugenio Astigarraga Echeverría <sup>1</sup> \*  
Arantza Mongelos García <sup>1</sup>  
Xavier Carrera Farran <sup>2</sup>

<sup>1</sup> Mondragon Unibertsitatea, España

<sup>2</sup> Universitat de Lleida, España

Partiendo de los cambios implementados en la Facultad de Humanidades y Ciencias de la Educación (HUHEZI) de Mondragon Unibertsitatea así como de los propiciados por el desarrollo del Espacio Europeo de Educación Superior (EEES), en este artículo se presenta una experiencia de innovación educativa -en cuanto a redefinición curricular, metodológica y evaluativa- desarrollada en los Grados de Educación Infantil y Educación Primaria de esta Facultad en los cursos 2017-18 (1º) y 2018-19 (1º y 2º) con una participación de 12 tutores, 40 profesores y 365 alumnos. Se fundamenta cómo, tras los cambios curricular y metodológicos habidos en la universidad, es necesario replantear el sistema de evaluación, acorde con dichas transformaciones y con las tendencias actuales de evaluación en educación superior. Se detalla la estrategia de evaluación diseñada a partir del perfil profesional y de las competencias a desarrollar en ambos grados y que se articula alrededor de los Resultados de Aprendizaje. Se presenta, con detalle, cómo se implementa el nuevo modelo evaluativo en cuanto a los momentos en que se lleva a cabo, los procedimientos y acciones requeridas, los agentes implicados, los instrumentos utilizados, las interacciones y feedbacks proporcionados, y los criterios evaluativos y de obtención de calificaciones establecidos. Finalmente, y sin intención de obtener generalizaciones -al tratarse de una experiencia que está en proceso de desarrollo-, se reflexiona sobre los aspectos y temáticas clave que pueden facilitar -o bien dificultar- el asentamiento y consolidación de las transformaciones que este tipo de innovación impulsan.

**Palabras Clave:** Evaluación; Innovación educativa; Resultados de aprendizaje; Universidad.

Based on the changes implemented at the Faculty of Humanities and Education Sciences (HUHEZI) in Mondragon Unibertsitatea, as well as those promoted by the development of the European Higher Education Area (EHEA), an experience of educational innovation - in terms of curricular, methodological and evaluative redefinition - is presented. The present experience was developed in the Degrees of Infant Education and Primary Education in this Faculty during the academic years 2017-18 (Year 1) and 2018-19 (Years 1 and 2) with a participation of 12 tutors, 40 teachers and 365 students. It shows how, after the curricular and methodological changes implemented in the university, it is necessary to rethink the evaluation system, in accordance with those transformations and with the current trends of evaluation in higher education. The evaluation strategy described, based on the professional profile and competencies to be developed in both degrees, is designed and articulated around the corresponding learning outcomes. It is presented, in detail, how the new evaluation model is implemented in terms of when it is carried out, the procedures and actions required, the agents involved, the instruments used, the interactions and feedback given, together with the evaluation and qualification criteria established. Finally, and without aiming to seek generalizations - since this is a particular experience that is still in process of development - we seek to reflect on the key aspects and themes that can facilitate -or hinder- the settlement and consolidation of the transformations that this type of innovation drives.

**Keywords:** Assessment; Educational innovation; Learning outcomes; University.

---

\*Contacto: eastigarraga@mondragon.edu

issn: 1989-0397  
www.rinace.net/riee/  
https://revistas.uam.es/riee

Recibido: 3 de septiembre de 2019  
1ª Evaluación: 8 de octubre de 2019  
2ª Evaluación: 15 de noviembre de 2019  
Aceptado: 21 de noviembre de 2019

## 1. Introducción

Teniendo como antecedentes -a nivel local- el Proyecto Mendeberry y -a nivel internacional- la construcción del Espacio Europeo de Educación Superior, iniciamos en la Facultad de Humanidades y Ciencias de la Educación de Mondragón Unibertsitatea<sup>1</sup> un proceso de innovación curricular que, de forma resumida, puede consultarse en Ozaeta, Mongelos, Astigarraga y Garro (2018). Este proceso de rediseño de un nuevo currículum, se implementó el curso 2017-18 en el primer curso tanto del Grado de Educación Primaria como del Grado de Educación Infantil, y en el curso 2018-19 se ha extendido al segundo curso de ambos Grados de Educación.

El modelo competencial sobre el que se han desarrollado los diferentes trabajos ha buscado ser coherente con el Marco Europeo de Cualificaciones (EQF-MEC) y su adaptación al ámbito de la Educación Superior en el estado español mediante el Marco Español de Cualificaciones para la Educación Superior (MECES), que se estableció a través del Real Decreto 1027/2011, y fue posteriormente modificado mediante el Real Decreto 96/2014. A través del establecimiento de los diferentes Marcos de Cualificaciones (CEDEFOP, 2018) el modelo de formación basada en competencias toma como referente para su diseño, desarrollo y evaluación los Resultados de Aprendizaje -en adelante RA- (ANECA, 2013; CEDEFOP, 2017; CEDEFOP, 2019; Comisión Europea, 2009) modelo que en el estado español es apreciable de forma bien asentada en la Formación Profesional (Astigarraga y Carrera, 2018).

El cambio curricular realizado en HUHEZI parte de revisar y ajustar el perfil de salida del alumno tomando como referentes las principales Funciones Profesionales y Competencias asociadas a las mismas. En base a este reajuste se establecen los RA<sup>2</sup>, que son la guía para la definición de las actividades de aula (Ozaeta et al., 2018). Toda esta transformación, dirigida desde un grupo dinamizador, se desarrolla con la participación abierta del conjunto del profesorado implicado en los respectivos Grados de Educación. El producto resultante -RA para el Grado en su conjunto y para cada uno de los cursos del mismo (véase figura 1) - requiere tiempo y trabajo colaborativo, que además de la obtención del propio producto, tiene como finalidad la mejora y el crecimiento profesional de todos los docentes de la Facultad, a la par que se orienta hacia la mejora de la actividad docente en la línea que señala Hattie (2017) “la planificación conjunta de las lecciones es la tarea con una de las más altas probabilidades de establecer una marcada diferencia positiva en el aprendizaje del alumno” (p. 95). El pasar del yo al nosotros, del trabajo individual/aislado del docente al trabajo en equipo de los docentes, es una tendencia que es cada vez más requerida (Guerriero, 2017; Paniagua e Istance, 2018), y que en nuestro contexto está resultando efectiva y clave para el cambio y la innovación en las instituciones

---

<sup>1</sup> En adelante HUHEZI, por sus siglas en euskera (lengua vehicular y de comunicación de esta Facultad).

<sup>2</sup> Desde el CEDEFOP (2017: 29) se diferencia -si bien se subraya su interrelación- entre Resultados de Aprendizaje Pretendidos (*Intended Learning Outcomes*) y Resultados de Aprendizaje Logrados (*Achieved Learning Outcomes*); a nuestros efectos, en la planificación nos referimos, obviamente, a los primeros.

educativas, tal como se evidencia a partir de la investigación realizada el presente año por Sarobe, López-Salas y Astigarraga (2019).

COMPETENCIAS		Resultados de Aprendizaje (Grado Completo)	Resultados de Aprendizaje (1º)	Resultados de Aprendizaje (2º)	Resultados de Aprendizaje (3º)	Resultados de Aprendizaje (4º)
Diseñar, desarrollar y evaluar procesos de enseñanza-aprendizaje y contextos saludables (espacios, recursos, relaciones, comunicación, metodologías, agrupaciones...), que respondan a la diversidad de los alumnos, impulsando el desarrollo, el aprendizaje, la participación y la convivencia de los mismos.	Conocer los conceptos básicos para impulsar el desarrollo físico, cognitivo y socioemocional de los alumnos; identificar estrategias y desarrollar y evaluar intervenciones efectivas.	Identificar las características de desarrollo cognitivo y socio-afectivo de los niños de 6 a 12 años, explicando la influencia de la escuela en dichos desarrollos.	Identificar y aplicar herramientas para la caracterización del desarrollo físico, cognitivo y socioafectivo de los alumnos, reflexionando sobre su impacto.	Desarrollar y aplicar en el aula estrategias para el desarrollo físico, cognitivo y socioemocional de los alumnos, compartiendo sus resultados con el grupo.	Tomar en cuenta las características del desarrollo físico, cognitivo y socioemocional de los niños de 6 a 12 años a la hora de diseñar contextos de enseñanza-aprendizaje, y desarrollar y evaluar intervenciones efectivas para promover su desarrollo saludable, extrayendo conclusiones de la reflexión sobre todo ello.	
Comunicarse de forma apropiada y efectiva en una variedad de situaciones comunicativas en la escuela, con actitud responsable, teniendo en cuenta los aspectos sociolingüísticos y asegurando el multilingüismo.	Producir textos oralmente o por escrito (en euskera y español) en el nivel C1, en inglés en el nivel B2 - en la mención de LE en el nivel C1) para responder adecuadamente a situaciones comunicativas reales.	Producir textos orales y escritos de manera correcta para comunicarse en situaciones comunicativas reales (en euskera y español), al menos, en el nivel B2, y en inglés en el nivel B1.	Producir textos orales y escritos para responder correctamente a situaciones comunicativas reales (en euskera y español), al menos, en el nivel B2, y en inglés en el nivel B1 - en la mención de LE, en el nivel B2).	Producir textos orales y escritos para responder correctamente a situaciones comunicativas reales (en euskera y español), al menos, en el nivel C1, y en inglés en el nivel B2 - en la mención de LE, en el nivel C1).	Producir textos orales y escritos para responder correctamente a situaciones comunicativas reales (en euskera y español), al menos, en el nivel C1, y en inglés en el nivel B2 - en la mención de LE, en el nivel C1).	
Identificar de forma crítica la influencia de la ciencia y la tecnología en nuestra sociedad y en el medio ambiente, participando de forma activa en minimizar sus impactos negativos.	Analizar de forma crítica la información relacionada con la ciencia y la tecnología, teniendo en cuenta la presencia de las mismas en los diferentes aspectos de la vida, valorando las posibilidades y dificultades para llevarlas al aula y buscar propuestas o alternativas a sus usos e impactos (ecológico, social y económico).	Reconocer la presencia de la ciencia y la tecnología en diferentes aspectos de la vida, valorando su impacto a fin de ir generando una conciencia crítica.	Identificar propuestas y estrategias para llevar al aula información y temáticas relacionadas con ciencia, tecnología y sociedad, tomando en cuenta la importancia de su presencia en los diferentes ámbitos de la vida y reflexionando sobre sus implicaciones y consecuencias.	Diseñar y desarrollar propuestas y estrategias para llevar a aula información y temáticas relacionadas con ciencia, tecnología y sociedad, tomando en cuenta la importancia de su presencia en los diferentes ámbitos de la vida y reflexionando sobre las implicaciones que sus distintas aplicaciones pueden tener.	Analizar las posibilidades que hay para desarrollar desde una perspectiva globalizadora propuestas educativas en Educación Primaria, tomando en cuenta la importancia de la ciencia y la tecnología en los diferentes ámbitos de la vida.	
Identificar y aplicar de forma crítica los documentos vigentes en el ámbito educativo, y, en particular, los que desarrollan el trabajo docente (PEC, PAG, DCB...), trabajando colaborativamente con otros profesores y agentes educativos.	Conocer documentos relacionados con la educación e identificar, analizar y evaluar de manera crítica su aplicación curricular, valorando el contexto, sus bases, sus implicaciones... a fin de analizar y valorar críticamente la evolución histórica del sistema educativo.	Conocer la ley de educación actual, las opiniones y propuestas de los agentes educativos, así como los DCB correspondientes, identificando los agentes educativos más importantes y el trabajo que realizan.	Identificar las leyes, reglamentos y documentos más relevantes en materia educativa que afectan a las escuelas, reflexionando sobre el impacto y los efectos de su aplicación en las aulas.	Identificar las leyes, reglamentos y documentos más relevantes en materia educativa que afectan a las escuelas, reflexionando sobre el impacto y los efectos de su aplicación en el centro educativo.	Diseñar, analizar el impacto y evaluar documentos para su aplicación en el centro, tomando en cuenta la aplicación curricular de las diferentes leyes, normativas y documentos que tienen incidencia en la escuela, extrayendo las conclusiones pertinentes. Desarrollar diversos elementos utilizados en la gestión escolar: informes escritos, elaborar planificaciones (cronogramas, hacer defensas orales...	

Figura 1. Ejemplos de Secuencia de: Competencia – Resultados de Aprendizaje para el Grado – Resultados de Aprendizaje por Cursos  
Fuente: Elaboración propia.

De esta manera, constatamos que si bien el trabajo participativo y colaborativo en el ámbito curricular puede iniciarse desde alguno de sus componentes, debemos anticipar que va a afectar al conjunto del ecosistema que es la institución educativa. En este sentido es clarificadora la postura de Scott (2015a) cuando señala (mencionando a Trilling y Fadel, 2009, p. 115) que “investigaciones recientes indican que para normalizar el aprendizaje colaborativo será necesario introducir cambios en los planes de estudios, la docencia, las prácticas de evaluación, los entornos de aprendizaje y el desarrollo profesional de las y los docentes” (p. 5). Esta misma autora, en otro de sus trabajos para la UNESCO (Scott, 2015b) nos remite a la necesidad de ampliar la perspectiva de la evaluación en los procesos de enseñanza-aprendizaje.

## 2. Fundamentación: del cambio metodológico al cambio en la evaluación

La perspectiva de la evaluación –en función del modelo/paradigma educativo en el que explícita o implícitamente se situaba el profesorado- ha ido cambiando a lo largo del tiempo, y, a menudo, se constituye en tema de amplios debates y difíciles acuerdos.

Desde finales de los años 80 del pasado siglo (Frey, Schmitt y Allen, 2012), se fue extendiendo la denominada evaluación auténtica, siendo su principal objetivo mejorar el aprendizaje de los alumnos implicándolos en los procesos de evaluación (Brown, 2015), para lo que debía cumplir con el principio de alineamiento constructivo de Biggs (2014) entre RA, actividades de evaluación y actividades de aprendizaje (figura 2).

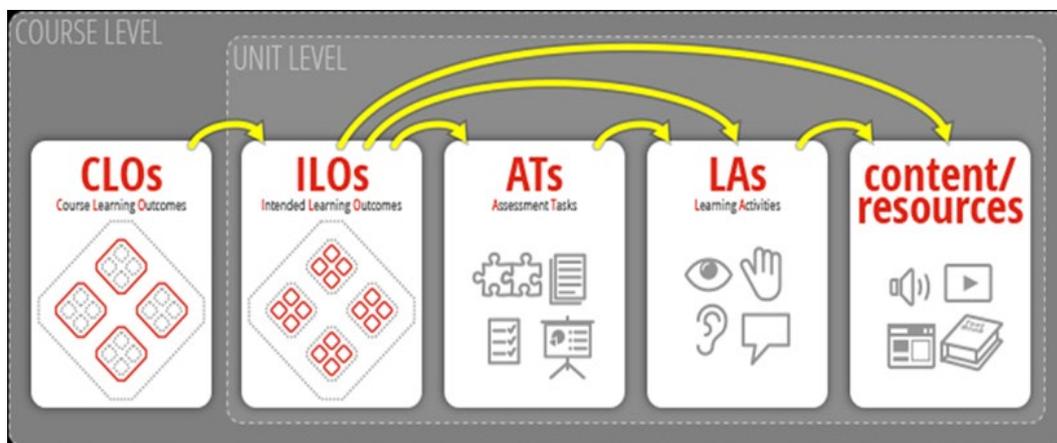


Figura 2. Diagrama del Alineamiento Constructivo (Rundle y Gurney, 2017)  
Fuente: Elaboración propia.

En este contexto, Brown (2015) establece entre las principales características de la evaluación auténtica, las siguientes:

- Orientada a la acción en diversidad de contextos relevantes;
- Respalda por evidencia relevante surgida de la práctica profesional;
- Matizada, articulada y transparente en la toma de decisiones;
- Verdaderamente representativa del esfuerzo y del rendimiento del estudiante;

- Con criterios ponderados en función de su importancia en términos de los RA;
- Proclive a maximizar el esfuerzo y el tiempo dedicados a la tarea por parte de los docentes;
- Incluyente en sus planteamientos, de modo que permita que todos los estudiantes lleguen al máximo de lo que son capaces.

SantaCruz (2019), siguiendo a Biggs y Tang (2011), relaciona el alineamiento constructivo con el aprendizaje profundo (deep learning), al tiempo que indica que la puesta en práctica de un proceso de aprendizaje superficial (surface learning) es un indicador de la falta de alineamiento constructivo entre los tres elementos -RA, tareas de evaluación, actividades de aprendizaje- que intervienen en el mismo (figura 2), y concluye que si bien “la transformación e innovación educativa no es una tarea fácil de lograr, el marco de alineamiento constructivo ayuda a los educadores docentes a reflexionar sobre sus propios diseños e implementaciones de enseñanza y aprendizaje” (SantaCruz, 2019, p. 7).

En este proceso reflexivo es importante una correcta identificación y definición de los RA, que indicarán a los estudiantes qué es lo que se considera importante de aprender, así como la secuencia y progresión que dichos aprendizajes van a seguir. Este trabajo de identificación y concreción de los RA requiere, tal como afirman Ozaeta et al. (2018), de la participación activa del conjunto del profesorado, situando dicha tarea en un marco de cambios educativos que acertadamente reflejan Siarova, Sternadel y Mašidlauskaitė (2017) al indicar que: Los objetivos de aprendizaje ya no están exclusivamente asociados a habilidades relacionadas con una determinada disciplina académica. Se espera que la educación “desarrolle las competencias de los individuos’ para enfrentar problemas y demandas complejas, movilizandolos recursos psicosociales, conocimientos, habilidades y actitudes adquiridas previamente en situaciones de aprendizaje similares a los contextos que se les presentarán en su vida diaria, profesional o académica” (p. 27).

Así pues, los RA, además de señalar los objetivos a desarrollar por los equipos docentes, ofrecen pistas de gran interés en relación con los contenidos a desarrollar -asociados a los distintos tipos de competencias-, con las formas de trabajo en el aula -pertinentes y válidas para el desarrollo y logro de todo ello-, así como con los espacios o entornos de aprendizaje en los que los procesos educativos deben desarrollarse -lo más parecidos y cercanos posibles a los contextos de aplicación de las competencias-.

Por otra parte, la ampliación de las competencias a desarrollar -incluyendo, junto a las competencias técnicas, las competencias del siglo XXI- genera dificultades en los procesos evaluativos; tal como señalan Siarova et al. (2017), “la evaluación de competencias clave y competencias transversales es un reto, ya que se refieren a constructos complejos que no son fácilmente medibles”. Y añaden que “la evaluación de las competencias sociales y emocionales de los alumnos se realiza, generalmente, desde una perspectiva formativa... que todavía es menos transparente que la evaluación de los logros académicos” (p. 29).

Para Hill y Barber (2014) es necesario ajustar la evaluación como parte del proceso continuo de repensar el aprendizaje y la enseñanza. Esta necesidad de repensar las formas de evaluación, para alinearlas con las actividades, los objetivos, las dinámicas de aula... es también un punto esencial del Proyecto Mendeberry 2025 en el que se subraya que “la evaluación es un tema central en cualquier cambio de paradigma educativo ya que está comprobado que difícilmente cambian los alumnos su forma de aprender si los

aprendizajes conseguidos se evalúan en función de modelos evaluativos previos que no cambian” (García, Zubizarreta y Astigarraga, 2017, p. 51).

Si bien en estas últimas décadas se han ido extendiendo propuestas centradas en el enfoque de evaluación auténtica, el tema de la evaluación sigue siendo un amplio campo de controversia y de dificultades innegables (Pellegrino, 2017; Siarova et al., 2017) en el que se van consolidando algunas tendencias y prácticas de la evaluación que deberán ir contrastándose a lo largo del tiempo. En este sentido son clarificadoras las palabras de Pellegrino (2017) cuando afirma que, “si bien se ha progresado en la evaluación de las habilidades cognitivas, se necesita mucha más investigación para desarrollar evaluaciones de las habilidades interpersonales e intrapersonales que sean adecuadas para los usos formativos y sumativos de la evaluación en entornos educativos” (p. 245).

En síntesis, las tendencias actuales en la evaluación de las competencias en la Educación Superior requieren que el abordaje de la evaluación se haga desde múltiples miradas (sentido y funciones de la educación, metodologías, dinámicas de aula, rol de docentes y alumnos, objetivos...) lo que conlleva una visión poliédrica, sistémica y global de los procesos educativo-formativos que vamos a implementar con enfoques innovadores, así como de la sociedad en la que se desarrollan dichos procesos.

### ***2.1. Características de un nuevo marco para la evaluación en la universidad***

La definición y concreción de un marco para la evaluación, conlleva responder -de forma actualizada, integrada y pertinente- a preguntas ya clásicas, que podemos resumir en: ¿para qué evaluar?, ¿qué evaluar?, ¿cómo y cuándo evaluar?, ¿quién evalúa?, ¿qué uso se hace de la información obtenida en la evaluación? ...

Las respuestas a las preguntas anteriores pueden tener muy diversas respuestas en función del foco y finalidad(es) de la evaluación que se asuman y de los destinatarios de la formación. En el caso de la evaluación de los procesos de enseñanza-aprendizaje que se desarrollan en los Grados de Educación Infantil y Primaria, una de las funciones principales de la evaluación es la acreditación del logro de los mínimos exigidos para el desempeño docente.

Ahora bien, las tendencias señaladas anteriormente, la demanda de que los nuevos perfiles profesionales incorporen las competencias del siglo XXI -que no son exclusivas de esta profesión, ni tan siquiera del ámbito académico o laboral-, las organizaciones internacionales (OCDE, UNESCO, WEF), el propio proyecto Mendeberry 2025... nos obligan a plantearnos una evaluación que va más allá de la certificación. Es más, junto a la extendida idea de que el alumno ha de ser responsable de su aprendizaje (a lo largo de la vida), se va afianzando -como complemento imprescindible de/para la misma- que el alumno ha de asumir la responsabilidad de su evaluación.

En este sentido, constatamos que -actualmente (Lee, 2013; SantaCruz, 2019; Siarova et al., 2017)- los tres principales enfoques de la evaluación -no excluyentes entre sí- podemos mencionarlos (figura 3) como:

- Evaluación del aprendizaje (Assessment of learning)
- Evaluación para el aprendizaje (Assessment for learning)
- Evaluación como aprendizaje (Assessment as learning)

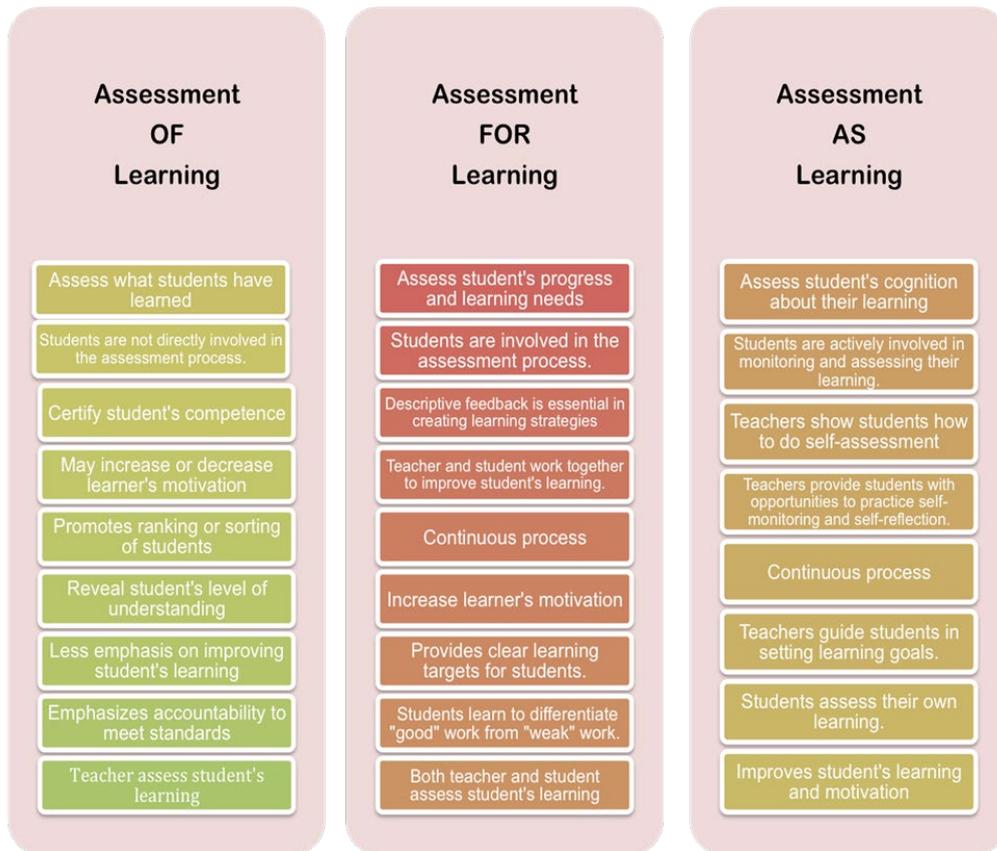


Figura 3. Evaluación del / para el / como aprendizaje (Lee, 2013)

Fuente: Elaboración propia.

El primero de estos enfoques se relaciona fácilmente con la evaluación sumativa y con la perspectiva certificadora de los procesos de enseñanza-aprendizaje. El segundo enfoque se puede asimilar -desde una perspectiva histórica- a los procesos de evaluación diagnóstica y, más específicamente, a la evaluación formativa. La combinación en la práctica de ambas formas de actuar ya conlleva cambios notables sobre los procesos de evaluación tradicionales que se han ido realizando en el ámbito universitario o en la Formación Profesional. En este último ámbito, ya se viene impulsando desde hace unos años este cambio hacia lo que se viene denominando evaluación como evolución (Tkніка, 2019):

*El enfoque Ethazi del proceso de evaluación de los alumnos es un enfoque basado en la evolución del grado de adquisición de sus competencias tanto técnicas como transversales, a medida que las ejercita en los sucesivos retos en los que trabajan con sus compañeros [...] La evaluación es un proceso que requiere de la participación de todos los actores: profesores y alumnos, principalmente, pero también personas del exterior (expertos, tutores de empresa...).* (p. 4)

Cuando se estima oportuno, se desarrolla un proceso de evaluación a 360 grados donde tanto profesores como alumnos y profesionales pueden opinar sobre el grado de adquisición de las competencias de cada alumno. Fruto de estas evaluaciones multi-enfoque se va ofreciendo -a lo largo del proceso- el feedback necesario, que posibilita la mejora de los procesos que se están desarrollando en el contexto del reto. También con ello, se obtienen unos resultados que son ofrecidos a cada alumno y/o a cada equipo en un formato que invite a la reflexión, y son acompañados por un proceso de feedback (de final de reto) por parte de los profesores, cuyo objetivo es obtener unos compromisos y acciones de mejora por parte del alumno/equipo de cara a la ejecución de los siguientes retos (p. 4).

El tercero de los enfoques -también en un contexto de evaluación formativa, pero, orientada así mismo a constatar los logros adquiridos-, permite facilitar al alumno el proceso de aprendizaje sobre su propia evaluación, de manera que el aprender a aprender se desarrolle también en relación con la evaluación de su propio proceso formativo y del logro de sus aprendizajes. Este proceso de metacognición debe ser facilitado y guiado por los docentes, los tutores... en definitiva, en el ámbito académico, por todos aquellos que participan en el proceso educativo-formativo del alumnado.

Con todo, esta aproximación por partes a la evaluación no debe perder su carácter de visión conjunta orientada al aprendizaje. Tal como señala Hayward (2015):

*Las preposiciones que vinculan la evaluación con el aprendizaje, como, para y de, pueden ser útiles si enfocan la atención en diferentes propósitos para la evaluación. Sin embargo, existe el peligro de que estas preposiciones se conviertan en un mantra irreflexivo que distraiga la atención del constructo clave: la evaluación es el aprendizaje. Centrarse en la evaluación para el aprendizaje en las aulas es una condición necesaria pero no suficiente [...] debemos reconocer la evaluación como aprendizaje: las preposiciones, por lo tanto, pueden reflejar diferentes propósitos de evaluación, pero la razón de ser de la evaluación es el aprendizaje.*  
(p. 38)

La puesta en práctica de lo mencionado anteriormente requiere disponer de momentos y estrategias que permitan desarrollar todo ello, pero, previamente, demanda identificar el qué se va a evaluar. Tal como se ha indicado en la introducción, nuestro referente de evaluación son los RA (Astigarraga y Carrera, 2018) derivados de las competencias identificadas (Ozaeta et al., 2018) previamente. El trabajo sobre los RA se ve facilitado por el uso de las taxonomías existentes en educación. En nuestro contexto son diversas las que podemos encontrar (Dreyfus, Anderson y Krathwohl, Marzano y Kendall...) siendo la más conocida y utilizada la taxonomía de Bloom. Desde nuestra perspectiva, y en consonancia con el enfoque de alineamiento constructivo y del contexto en que se desarrolla esta experiencia de innovación, se opta por utilizar la taxonomía SOLO de Biggs (2011, 2014).

### **3. Contexto de la experiencia**

La Facultad de Humanidades y Ciencias de la Educación (HUHEZI) de Mondragón Unibertsitatea inició en el curso 2014-2015 (figura 4) un proceso de cambio que toma como eje el desarrollo de un currículum basado en competencias y operativizado por medio de "Propuestas de Trabajo" inspiradas en retos con una perspectiva inter/transdisciplinar (Ozaeta et al., 2018).

Este proceso de innovación educativa se inicia en el Grado de Comunicación Audiovisual, y se le da continuación en los grados de Educación Infantil y Educación Primaria. La experiencia de la Facultad en el desarrollo del Proyecto Mendeberry, y, posteriormente, en la adecuación de las titulaciones al Espacio Europeo de Educación Superior, sirve como referencia previa para avanzar en la definición y la aplicación de un nuevo modelo educativo más ajustado a las necesidades emergentes que presenta este siglo XXI.

El curso 2017-18 se inicia la aplicación de este nuevo modelo formativo y evaluativo participando en el mismo 2 grupos del Grado de Educación Infantil y 4 grupos del Grado de Educación Primaria; cada uno de los anteriores grupos tiene su tutor (6 tutores de curso) y participan -de forma parcial, en función de las distintas Propuestas de Trabajo- un total de 30 docentes y cerca de 180 alumnos.

Durante el curso 2018-19, han tomado parte en este proceso 4 grupos del Grado de Educación Infantil (2 de 1º y 2 de 2º) y 8 grupos del Grado de Educación Primaria (4 de 1º y 4 de 2º). Cada uno de los anteriores grupos tiene su tutor (12 tutores de curso) y han participado -de forma parcial, en función de las distintas Propuestas de Trabajo- un total de 30 docentes en 1er curso y cerca de 40 en el 2º curso. En cuanto al número de alumnos, han sido 190 alumnos en 1er curso y 175 en 2º curso.

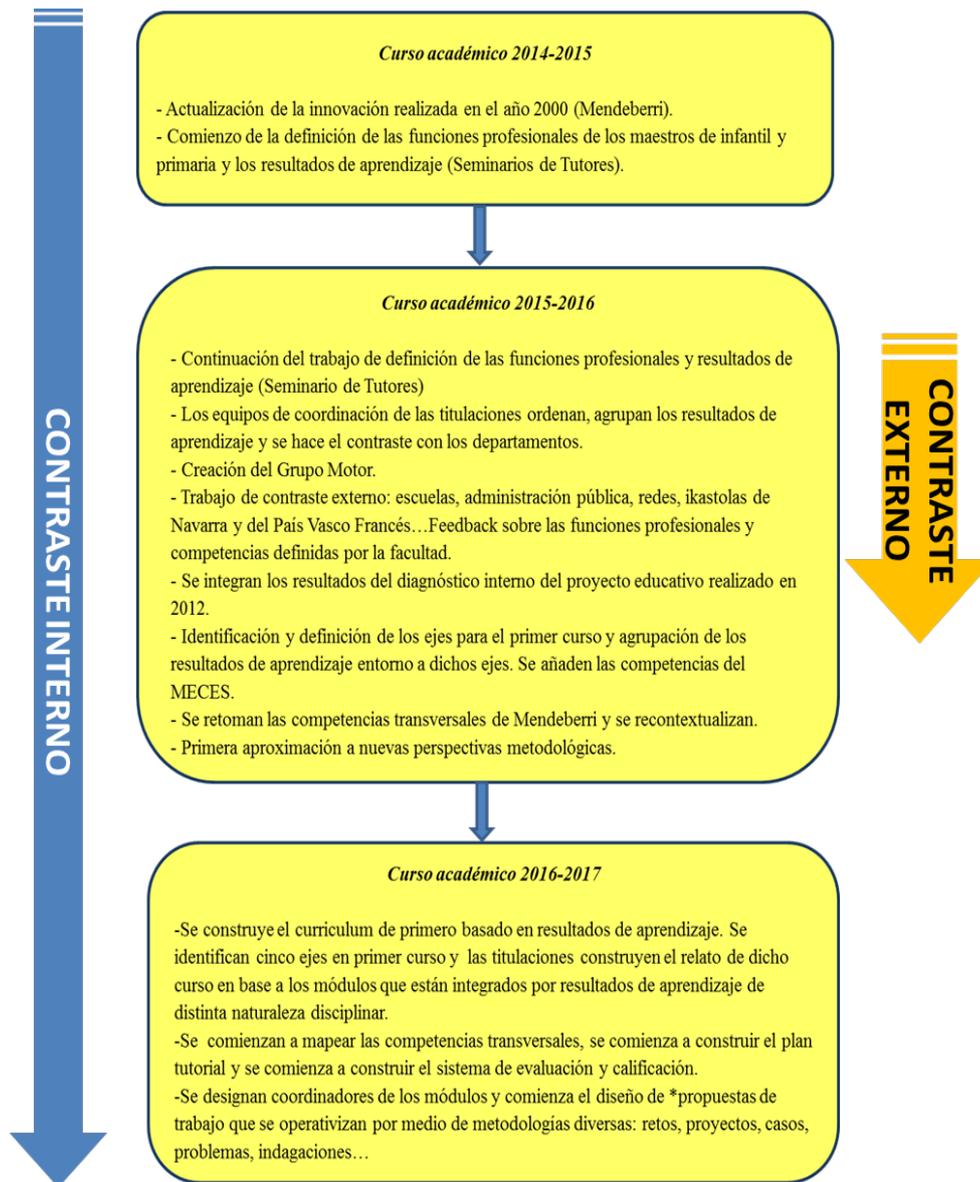


Figura 4. Proceso de rediseño curricular de los Grados de Educación Infantil y Educación Primaria  
Fuente: Elaboración propia.

## 4. Diseño de la estrategia de evaluación en base a los Resultados de Aprendizaje

En la perspectiva educativa que configura el marco para la evaluación definido en los puntos anteriores, la estrategia de evaluación se va definiendo a la par que se estructura la propuesta curricular para el aula. Tras haber llegado a la concreción de los RA (Ozaeta et al., 2018), se comienza a desplegar el proceso que se muestra de forma esquemática en la figura 5.

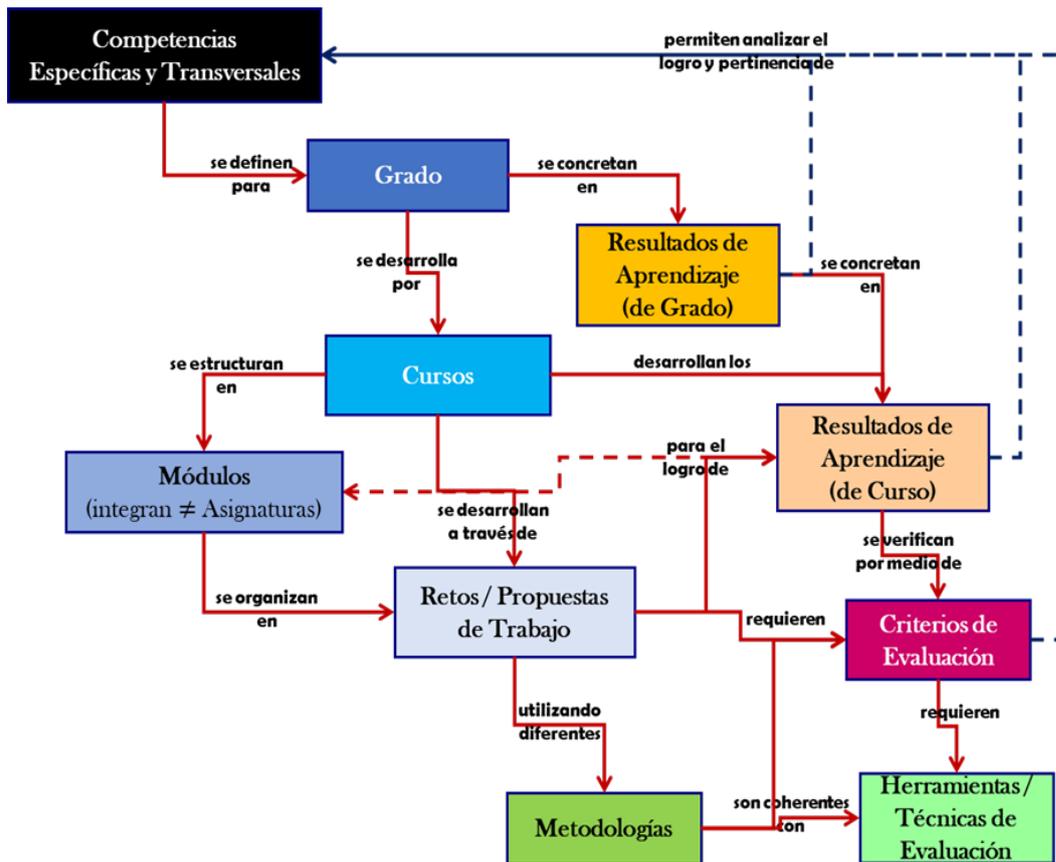


Figura 5. Proceso para el diseño curricular en base a los Resultados de Aprendizaje  
Fuente: Elaboración propia.

De esta manera, se procede a un trabajo en paralelo de: a) revisión y ajuste de los RA del primer curso y b) definición y concreción de los Módulos sobre los que soportar tanto los procesos curriculares (temáticas y contenidos a trabajar, alcance y delimitación de los mismos...) como los procesos administrativos (unidad mínima de matrícula, criterios para superación/repetición de curso...).

La primera de estas tareas lleva al conjunto del profesorado a especificar los RA para cada uno de los cursos de acuerdo a los niveles de la taxonomía SOLO, asumiendo que al finalizar los procesos de enseñanza-aprendizaje no tendríamos ningún alumno o alumna en el nivel preestructural. Un ejemplo de ello puede verse en la figura 6. A su vez, conlleva establecer cuál será el mínimo a lograr en este cada curso. Normalmente, para cada RA

el mínimo establecido en los distintos casos ha sido el nivel 2 (multiestructural), o el nivel 3 (relacional).

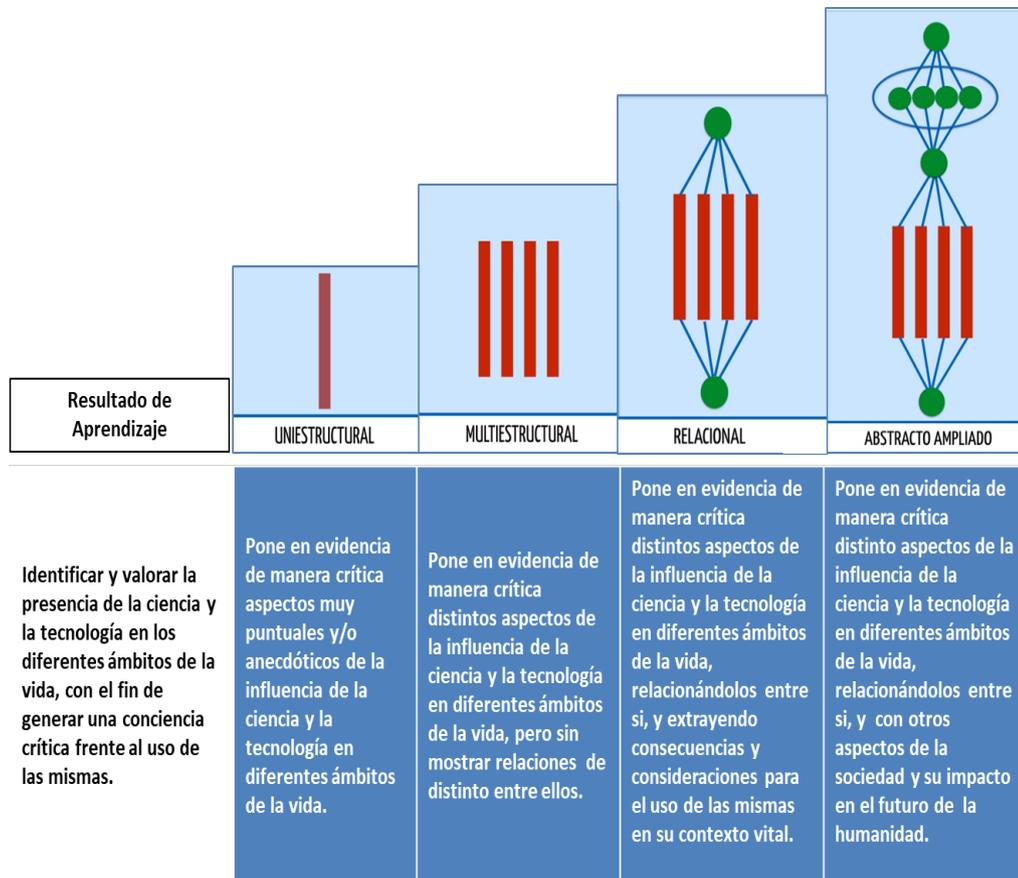


Figura 6. Ejemplo de nivelación de un Resultado de Aprendizaje  
Fuente: Elaboración propia.

Todo este proceso de definición del marco de evaluación, de comprensión de la taxonomía a utilizar, de especificación y nivelación de los RA... es complejo, pero asumimos que el mismo es imprescindible y permite una comprensión compartida de la evaluación por parte del conjunto del cuerpo docente. Tal como señalan Siarova et al. (2017) entre las condiciones “para garantizar prácticas de evaluación coherentes para el ‘aprendizaje del siglo XXI’, se encuentran:

- La importancia de una definición clara de las competencias de los estudiantes en términos de RA y su reflejo en el currículum;
- La formación docente, incluida la ITE [Initial Teacher Education], la etapa de inducción y la CPD [Continued Professional Development], que proporciona a los profesores una comprensión común de las competencias clave y la orientación de la evaluación a lo largo de sus carreras profesionales;
- Mecanismos de colaboración en forma de comunidades de enseñanza-aprendizaje (TLC Teaching and learning communities), con el objetivo de mejorar las prácticas de evaluación para alinearlas mejor con las competencias clave” (p. 45).

En el desarrollo de los pasos anteriores, se van compartiendo y asumiendo diferentes principios que dirigirán los procesos de evaluación posteriores. Entre estos principios, los más significativos son los siguientes:

- El foco debe estar en los RA, siendo la labor de los docentes facilitar e impulsar su desarrollo por parte de los alumnos.
- El referente sobre el que se trabaja son los Módulos<sup>3</sup>, por lo que las Propuestas de Trabajo definidas dentro de los mismos no se calificarán.
- La evaluación debe ser pensada desde el Módulo en su globalidad, haciéndose el desarrollo y seguimiento de los RA desde el mismo.
- Para la calificación del Módulo se tomará en consideración el peso porcentual asignado a cada RA en el mismo.
- La calificación de los RA se realizará al finalizar el Módulo. Para esta calificación se tomarán en consideración las calificaciones que se han otorgado a cada RA en cada uno de los distintos Módulos en los que está presente en la evaluación.
- Los alumnos no recibirán calificaciones de los RA hasta la finalización del curso. De esta manera, al finalizar el curso tendrán dos calificaciones<sup>4</sup>: la correspondiente a cada uno de los RA y, en función de ello, la obtenida en cada uno de los Módulos.
- Dado que hay RA que se desarrollan en más de un Módulo, se compartirán las rúbricas de nivelación de los mismos.
- Desde cada Módulo se diseña y realiza el seguimiento de los alumnos en base a las correspondientes rúbricas de RA definidas previamente.

Continuando con el proceso de diseño, un paso posterior es el de concretar los RA a desarrollar desde cada uno de los Módulos. Esta tarea requiere de una visión global del curso y de acuerdos entre los diferentes equipos docentes que se constituyen para cada Módulo, y conlleva también una primera aproximación a la ponderación de los RA, tanto para su calificación como para la calificación final de los Módulos. Una primera aproximación de este trabajo puede verse en la figura 7.

---

<sup>3</sup> Para el primer curso se definieron cinco Módulos. En el Grado de Educación Primaria fueron: I) La Escuela Inclusiva; II) Yo, futuro educador; III) Lengua, Pueblo y Escuela; IV) Ciencia y Tecnología en nuestra sociedad; V) Comunidad Educativa.

<sup>4</sup> En realidad, son tres calificaciones ya que, por cuestiones administrativas, debemos seguir otorgando calificaciones a las asignaturas oficiales del plan de estudios. Esto se realiza a través de una conversión de la calificación de los Módulos en calificaciones de asignaturas mediante la correspondiente ponderación (en función de la presencia de cada asignatura en los distintos Módulos).

EDUCACIÓN PRIMARIA		1. La Escuela Inclusiva	2. Yo futuro educador	3. Lengua, Pueblo y Escuela	4. Ciencia y Tecnología en nuestra sociedad	5. Comunidad Educativa
IE 1.1	Explicar la influencia de la escuela en el desarrollo del niño de 6-12 años, identificando las características de sus desarrollo cognitivo y socio-afectivo.	%40 / % 45	%20 / %55			
IE 1.2	Identificar las características de la escuela inclusiva, analizándolas y relacionándolas con los contextos que facilitan dar respuesta a la diversidad, desarrollando estrategias generales y específicas para dar respuesta a los niños que están en situaciones de desventaja.	%40 / %80	%10 / %20			
IE 1.3	Identificar las características, desarrollo y evolución de los distintos agentes educativos.	%5 / %80				%5 / %20
IE 1.4	Reflexionar y obtener conclusiones sobre el perfil profesional del maestro, partiendo de las experiencias personales.		%40 / %80		%5 / % 20	
IE 1.5	Identificar los principios y los actores de una Comunidad Educativa que aprende, reflexionando sobre sus consecuencias.	%10 / %100				
IE 1.6	Utilizar de manera crítica la tecnología y los medios audiovisuales, tanto para el desarrollo de los trabajos diarios como para fortalecer la comunicación.		%10 / %50		%10 / % 50	
IE 1.7	Producir textos/mensajes -escritos y orales- para dar respuesta adecuada a situaciones comunicativas reales (en euskera y español, al menos en el nivel B2, y en inglés en el nivel B1).		%20 / 40%	%10 / %30		%10 / %30
IE 1.8	Valorar la importancia del multilingüismo aditivo para seguir aprendiendo distintos idiomas, identificando las características, fortalezas y debilidades de cada cual.			%20 / %80		%5 / %20
IE 1.9	Identificar y caracterizar la importancia de ser agentes activos en el desarrollo de la comunidad vascoparlante, identificando las características, herramientas, marcos y recursos que presenta la situación lingüística de nuestro entorno y de la sociedad,			%60 / %80		%10 / %20
IE 1.10	Identificar la presencia de la ciencia y la tecnología en los diferentes aspectos de nuestras vidas, valorando su influencia en el desarrollo de una conciencia crítica.				%35 / %100	
IE 1.11	Identificar y valorar los principios básicos de la sostenibilidad comprendiendo los principales fenómenos que se dan en la sociedad actual, conociendo algunas experiencias en este ámbito para ir situándose en el marco de la transición socioecológica.				%35 / %100	
IE 2.1	Identificar la legislación educativa vigente, reconociendo a los principales agentes educativos y su trabajo, así como sus opiniones y propuestas, identificando los DCBs que de todo ello se derivan.					%5 / %100
IE 2.2	Analizar las características y claves del Sistema Educativo, y de los distintos tipos de redes, de Centros y de modelos existentes, identificando y valorando su evolución socio-histórica.			%10 / %20		%60 / %80
IE 2.3	Identificar los distintos cargos y órganos que hay en una escuela, reflexionando sobre su necesidad, importancia y calidad.	%5 / %100				
IE 2.4	Identificar distintos tipos de proyectos que tienen relación con el Centro, comprendiendo su sentido y finalidad.				%15 / % 100	

Figura 7. Relación entre Módulos y Resultados de Aprendizaje  
Fuente: Elaboración propia.

A continuación, desde cada Módulo se identifican posibles Propuestas de Trabajo (PT), que utilizando diferentes Metodologías (Resolución de Problemas, Estudio de Casos, Proyectos, Conferencias, Visitas...) permiten desarrollar los RA correspondientes a cada Módulo (figura 8). Cada Módulo dispone de una Guía del Alumno que presenta tanto los RA a desarrollar en el mismo como las distintas Propuestas de Trabajo que lo posibilitarán; así mismo, junto al equipo de docentes que se hará cargo del mismo, se da a conocer de forma genérica el proceso de evaluación que se seguirá. De manera similar, se desarrolla una Guía del Alumno para cada una de las Propuestas de Trabajo, en la que se presentan con una mayor concreción las tareas a desarrollar así como los RA que se van a trabajar en dicha PT y los Criterios de Evaluación relacionados con cada uno de los RA a desarrollar; al mismo tiempo, además del cronograma correspondiente, se explicitan los agentes y las formas en que los mismos van a participar en la evaluación, siendo también tarea del equipo de profesores el desarrollo de las necesarias herramientas de evaluación para cada una de las Propuestas de Trabajo (en relación con los correspondientes Criterios

de Evaluación<sup>5</sup>). Junto a estos procesos de evaluación, desde la Facultad, de manera genérica y periódica, se pasa también una encuesta de satisfacción (anónima) a los alumnos en relación con las PT desarrolladas.

2017-18	EDUCACIÓN PRIMARIA	Lunes	Martes	Miércoles	Jueves	Viernes
1 Semana	Septiembre 11-15	Mi personalidad: ¿cómo soy yo y por qué he venido aquí? (recepción/introducción)				
2 Semana	Septiembre 18-22					
3 Semana	Septiembre 25-29	¿Cómo se desarrolla el niño de 6 a 12 años?				
4 Semana	Octubre 2-6					
5 Semana	Octubre 9-13					
6 Semana	Octubre 16-20	La ciencia en la vida y en la escuela				
7 Semana	Octubre 23-27	¿Qué educador seré / quiero ser?				
8 Semana	Octubre 30 - Noviembre 3					
9 Semana	Noviembre 6-10	Comunidad - ¿en qué comunidad vivimos?				
10 Semana	Noviembre 13-17					
11 Semana	Noviembre 20-24	Comunidad Educativa + Evolución de la Escuela (evolución socio-histórica)				
12 Semana	Noviembre 27 - Diciembre 1	Comunidad Educativa + Evolución de la Escuela + Video Project				
13 Semana	Diciembre 4-8					
14 Semana	Diciembre 11-15	Construyendo el perfil del educador				
15 Semana	Diciembre 18-22					
	Navidad					
16 Semana	Enero 8-12	El polvo de las estrellas como materia de vida				
17 Semana	Enero 15-19					
18 Semana	Enero 22-26					
19 Semana	Enero 29 - Febrero 2	¿Cómo son los contextos educativos que posibilitan el éxito de todos los niños?				
20 Semana	Febrero 5-9	Yo y la diversidad (etiquetas, prejuicios, estereotipos) - lo que yo he hecho y lo que a mí me ha tocado sufrir + historias de vida				
21 Semana	Febrero 12-16					
22 Semana	Febrero 19-23	Colectivos y redes, agentes...				
23 Semana	Febrero 26 - Marzo 2	Diversidad de contextos sociolingüísticos: comunidad educativa + profesor				
24 Semana	Marzo 5-9					
25 Semana	Marzo 12-16					
26 Semana	Marzo 19-23					
	Semana Santa					
27 Semana	Abril 9-13	(modelos educativos y lingüísticos, perfil del profesor...)				
28 Semana	Abril 16-20	Diversidad de contextos sociolingüísticos: comunidad educativa + profesor (modelos educativos y lingüísticos, perfil del profesor...)				
29 Semana	Abril 23-27	¿Qué tipo de propuestas debe hacer el profesor de EP para ofrecer una educación de calidad a todos los alumnos?				
30 Semana	Abril 30 - Mayo 4					
31 Semana	Mayo 7-11	¿Conocemos el medio ambiente del País Vasco?				
32 Semana	Mayo 14-18					
33 Semana	Mayo 21-25	Evolución de la sociedad: ¿qué es la transición socioecológica? ¿es necesaria?				
34 Semana	Mayo 28 - Junio 1	Más allá de nuestro sistema educativo				
35 Semana	Junio 4-8					
36 Semana	Junio 11-15	¿Qué recorrido he hecho en este curso? ¿Cuáles son los siguientes pasos?				
37 Semana	Junio 18-22					

La Escuela Inclusiva
Yo Educador
Lengua, Pueblo y Escuela
Ciencia y Tecnología en nuestra sociedad
Comunidad Educativa

Figura 8. Módulos (dcha.) y Propuestas de Trabajo de cada uno de ellos (2017-18)  
Fuente: Elaboración propia.

## 5. Desarrollo de la evaluación centrada en Resultados de Aprendizaje: el tutor como dinamizador

Pasar del diseño y de la planificación mostrados en el apartado anterior requiere, ante todo, de la coordinación ágil y eficiente de todos los docentes que participan en el desarrollo de los RA a lo largo del curso con un mismo grupo de alumnos en los distintos

<sup>5</sup> Como puede entender el lector experimentado, este relato “tan lineal y sencillo” en la práctica es mucho más complejo (necesidad de compartir visiones de la educación y la evaluación; conocer y compartir metodologías y formas de trabajo en el aula; identificar y consensuar funciones de la evaluación, así como estrategias y técnicas para ello; acordar la ordenación y secuencia de las distintas Propuestas de Trabajo...), y tampoco está finalizado para el primer día de comienzo del curso. Así pues, a menudo nos encontramos desarrollando una Propuestas de Trabajo y diseñando la(s) siguiente(s), por lo que es necesario subrayar que, además de trabajo y dedicación, hace falta una buena dirección, acompañamiento y ayuda a los docentes, así como una periódica revisión (normalmente a fin de curso) y mejora de los procesos desarrollados.

módulos y en un mismo módulo. Así mismo, y a lo largo del proceso, es de vital importancia el seguimiento realizado al alumnado por parte del tutor de curso a lo largo de todo el año. Ambos elementos se explican a continuación, incidiendo en los momentos, instrumentos y objetivos de cada uno de ellos.

Al comienzo del curso, a cada estudiante se le asigna un tutor, que a su vez es el docente responsable del módulo “Yo Educador” en el primer año del grado. Este módulo es el de mayor presencia a lo largo de todo el curso, para poder realizar un mayor seguimiento y acompañamiento al alumno; y es el que da inicio y con el que se finaliza el curso académico de primero.

El tutor realiza, al menos, una reunión por trimestre con todos los docentes que participan en las propuestas de trabajo de los módulos. El objetivo de dichas reuniones es, principalmente, realizar un seguimiento cualitativo de los alumnos del grupo de forma que el tutor pueda ir orientando a cada uno de sus estudiantes según sus necesidades. Los docentes comparten sus experiencias, así como sus inquietudes y preocupaciones tanto desde la perspectiva de evolución de los estudiantes, como en relación con el desarrollo de los diferentes módulos (y los RA asociados a los mismos).

Además de estas reuniones, a lo largo del curso se realizan otras dos sesiones de “evaluación” propiamente dichas. Una en enero-febrero y la otra al finalizar el curso, en junio (en la figura 8, quedan recogidas estas semanas sin denominación específica). La primera de ellas sirve para que el equipo docente, por un lado, y los estudiantes, por otro, valoren cómo van progresando en los RA hasta el momento. Para ello, y de manera previa a la reunión de evaluación, cada docente realiza la valoración de los RA trabajados hasta la fecha en su módulo, y aplicando las rúbricas descritas anteriormente, le asigna el nivel correspondiente tomando en consideración los criterios de evaluación establecidos en cada una de las Propuestas de Trabajo desarrolladas hasta ese momento. El instrumento de evaluación -las distintas rúbricas- es compartido por todo el profesorado. Cada docente desarrolla y concreta su evaluación, y en la reunión de evaluación, coordinada por el tutor, se llega a acuerdos para cada alumno cuando se dan diferencias en las asignaciones del nivel de logro en un RA presente en dos o más módulos. En esta primera evaluación, únicamente se valora al estudiante teniendo en cuenta los niveles definidos en las rúbricas, no se obtiene ninguna calificación.

Al mismo tiempo, cada estudiante realiza su propia auto-evaluación de los RA. Para ello, cuenta con una herramienta específica en la que además de indicar el nivel en el que se encuentra -teniendo en cuenta la rúbrica para cada RA-, el estudiante tiene que evidenciar y justificar la valoración que realiza. En esta primera valoración, puede darse el caso de que algunos RA aún no se hayan podido empezar a desarrollar y queden sin valorar.

Realizadas ambas valoraciones, el tutor/a tiene una sesión de seguimiento con cada estudiante, en la que comparte la valoración realizada por los docentes y la realizada por el propio estudiante, contrastándolas con los niveles mínimos establecidos para cada uno de los RA. Ello da como resultado un informe valorativo y una gráfica (figura 9) que sintetiza las diferentes valoraciones obtenidas (valoración del docente, mínimo estipulado, autoevaluación del alumno).

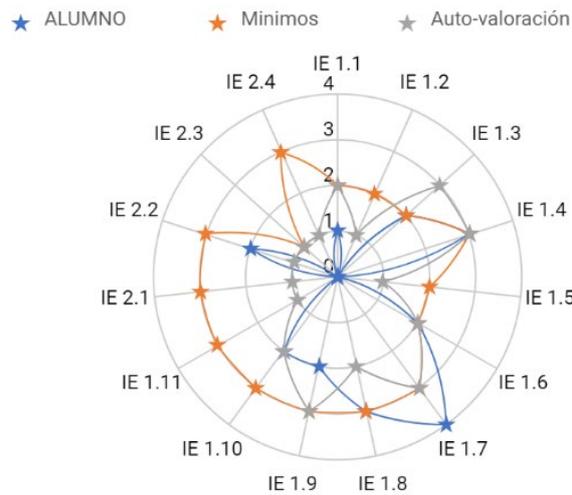


Figura 9. Síntesis gráfica de la primera valoración del progreso del alumno  
Fuente: Elaboración propia.

La sesión de tutoría tiene una duración de 30 minutos aproximadamente. El objetivo de la misma es compartir con cada estudiante la perspectiva del equipo docente con la que él/ella misma pueda tener de su progreso hasta el momento. En la misma sesión se acuerdan los objetivos de progreso para la segunda mitad de curso, analizando para ello las dificultades que se hayan podido detectar en cada caso, bien por parte del equipo docente bien por parte del propio alumno.

Al finalizar el curso, se realiza un proceso semejante de evaluación. En este caso, además de las valoraciones de cada uno de los RA asignados a cada módulo, los docentes han de calificar cada uno de dichos RA.

Para asignar las calificaciones, se han establecido unos baremos de forma que cada nivel tiene asignado un intervalo numérico, que permite diferenciar y precisar los logros de los alumnos aun cuando estén en el mismo nivel para un determinado RA, tal como se puede ver en la figura 10.

Mínimo = 1		Mínimo = 2		Mínimo = 3	
Nivel	Calificación	Nivel	Calificación	Nivel	Calificación
1	5,0-5,9	2	5,0 - 6,9	3	5,0-7,9
2	6-7,4	3	7,0 - 8,5	4	8,0-10,0
3	7,5-8,9	4	8,6 - 10,0		
4	9,0-10,0				

Figura 10. Baremación por niveles en RA  
Fuente: Elaboración propia.

La sesión de evaluación de junio, al igual que la primera de enero-febrero, sirve para acordar los niveles de consecución de cada uno de los RA, y en caso necesario adecuar la calificación. Se entiende que aun cuando el nivel de consecución del RA ha de ser el mismo en el caso de RA compartidos por distintos módulos, las calificaciones no tienen por qué

ser las mismas, siempre y cuando queden dentro de las franjas establecidas para cada nivel. Tras la evaluación, el tutor/a vuelve a tener una sesión de tutoría con cada alumno. En ella se contrastan las valoraciones realizadas por él mismo (mediante una herramienta similar a la utilizada en la sesión previa de enero, y en la que valora cada RA y justifica con evidencias dicha valoración) y las realizadas por el equipo docente con los mínimos establecidos. Los objetivos de esta tutoría son compartir con cada estudiante su situación con respecto a los RA a conseguir en el curso. En esta sesión el alumno, contrasta su progreso desde enero hasta junio (figura 11), posibilitando una aproximación a su calificación. Todos los alumnos acceden a las calificaciones una vez terminadas todas las tutorías a través de la secretaría virtual. Cada estudiante puede ver sus calificaciones para cada RA, así como para cada módulo, y finalmente, para cada asignatura.

En el transcurso de la sesión de tutoría, el estudiante y el tutor establecen las actividades a realizar para superar los RA no logrados, en caso de que los haya, así como objetivos de cara al siguiente curso en el que va a tener continuidad el desarrollo de los RA trabajados hasta el momento<sup>6</sup>. Estos compromisos quedan recogidos en su herramienta de evaluación y los retoma al comienzo del nuevo curso.

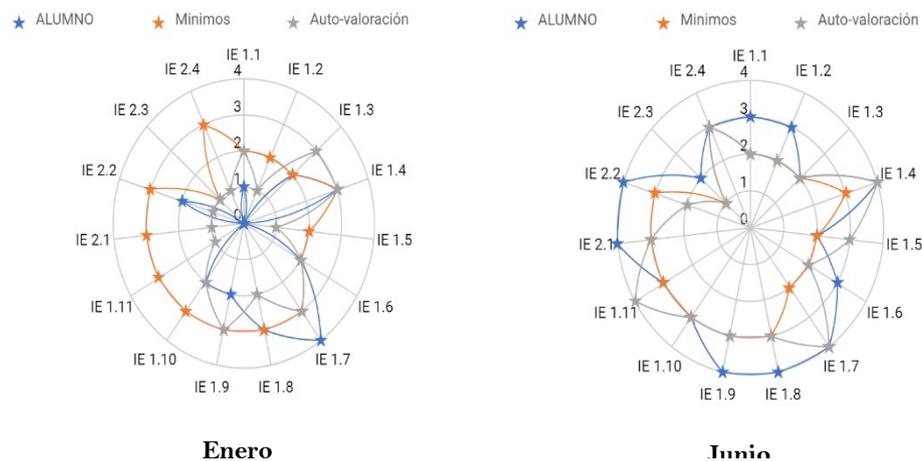


Figura 11. Síntesis gráfica de la progresión del alumno en su desarrollo de los RA – Enero (izda.) / Junio (dcha.)  
Fuente: Elaboración propia.

## 6. Limitaciones y dificultades observadas

Una reflexión crítica del desarrollo de la experiencia presentada nos lleva a clasificar las diferentes necesidades y dificultades que trae consigo el cambio curricular (objetivos, metodología, evaluación, rol docente...) en dos grandes ámbitos. El primero el relativo al diseño, conceptualización y visualización de los cambios que se quieren implementar. El segundo el relacionado con su puesta en práctica, seguimiento y consolidación.

<sup>6</sup> Como es lógico, puede haber algún RA nuevo, o bien puede darse que algún RA haya finalizado su proceso de desarrollo.

En relación con el primer ámbito, los aspectos más reseñables son los siguientes. En primer lugar, observamos que -en nuestro contexto, y por mucho que se verbalice o aparezca en los documentos administrativos- no hay una comprensión compartida de qué es la Formación Basada en Competencias (FBC), y, unido a ello, la orientación a los RA que, en el contexto anglosajón, es más habitual. Por tanto, una dificultad y una tarea inicial es la de compartir en qué consiste un modelo de FBC, cómo se estructura el mismo, qué tipo de competencias lo incluyen... León (2014), citando a Flores (2003), recoge bien la esencia de este tipo de formación al señalar que: Los modelos educativos por competencias están sustentados en tres ejes: la educación basada en competencias, la flexibilidad curricular y los procesos educativos centrados en el aprendizaje. Esto conduce a que “las prácticas educativas estén orientadas hacia la interdisciplinariedad, el trabajo grupal, el conocimiento aplicado a realidades concretas, el papel del docente como coordinador y facilitador del aprendizaje y la participación activa del estudiante en su proceso de formación” (p. 55).

Unido a lo anterior, otra gran dificultad que se nos ha presentado es doble: cómo identificar y describir los RA, a la par que diferenciamos los mismos de las competencias. Esta cuestión es importante ya que el describir adecuadamente los RA esperados debe permitir al equipo docente pensar más allá de la propia asignatura, reflexionar sobre los aportes que la misma hace al logro de dichos RA, al tiempo que facilita el diseño de Propuestas de Trabajo para el aula de carácter interdisciplinar y referenciadas en el contexto profesional para el que se está formando al grupo de alumnos.

Ya en su momento, también desde la ANECA (2013) se señalaban algunas de las dificultades y limitaciones que conllevaba el trabajar con los RA, y en particular subrayaba que: Definir el plan de estudios en términos de RA requiere dedicación, esfuerzo, recursos y obstáculos a sortear. El cambio del enfoque hacia un modelo centrado en el estudiante exige la concienciación del personal académico de las universidades, la familiarización con su uso y la dedicación de su tiempo y esfuerzo para reflexionar sobre los resultados que deben lograr los estudiantes, así como su trabajo en equipo para lograr objetivos comunes e integrados en un nivel superior. Supone, por tanto, “en muchas ocasiones una transformación significativa que a menudo tarda años en hacerse efectiva” (p. 18).

Otra dificultad se presenta a la hora de concretar la evaluación desde las diferentes Propuestas de Trabajo. Aquí también la dificultad es doble. Por una parte, describir Criterios de Evaluación relacionados con las tareas a desarrollar y orientados a la evaluación de los RA, es la esencia en la práctica del alineamiento constructivo... pero no es siempre tarea fácil. Por otra parte, que un mismo RA se trabaje desde Propuestas de Trabajo desarrolladas por equipos de profesores distintos, conlleva la necesidad de una comprensión compartida de dicho RA, a la par que se acota la misma, dando la posibilidad de que cada Propuesta de Trabajo tenga sus propios Criterios de Evaluación orientados a RA compartidos. Esto requiere mucha coordinación y seguimiento, a la vez que resulta imprescindible dar apoyo a los distintos equipos de docentes que están diseñando y desarrollando las diversas Propuestas de Trabajo.

En relación al segundo ámbito -centrado en su puesta en práctica, seguimiento y consolidación- la principal dificultad en la práctica es que el grupo de docentes responsables del curso (tutores y profesores) tengan una visión compartida tanto del significado y alcance de los RA como de las tareas/actividades, y consecuentes evidencias, que deben/pueden presentar los alumnos para justificar sus avances y progreso. La falta de tradición y cultura para trabajar con este enfoque -que busca superar el conocimiento

académico para centrarse en la introducción a la profesión y su progresivo desarrollo-, hace que todavía sean evidentes diferentes interpretaciones de los aportes de los RA, así como de las rúbricas y su aplicación.

Por otra parte, como puede deducirse del proceso explicado, se hace necesario disponer de tiempos y espacios para la coordinación, el diálogo y la negociación con el fin de consensuar los niveles de logro de los RA desde la mirada de las actividades/tareas realizadas en los distintos módulos. Evidentemente, el proceso se complica cuando para cada curso hay más de un grupo con su correspondiente equipo docente pues ello exige coherencia tanto al interior de cada uno de ellos (intra) como entre los distintos equipos docentes (inter).

Lo anterior requiere a su vez liderazgo y coordinación desde el equipo que dirige el proceso, a la par que fomenta una cultura de cambio educativo y ofrece suficientes espacios formativos para que el conjunto de docentes avance de manera coordinada en el marco propuesto, superando la cultura de la educación como transmisión y de la evaluación como calificación presente en parte de los docentes.

Desde la perspectiva de los estudiantes, la primera gran dificultad -paralela a la que evidencian los docentes noveles- es la de “deconstruir” el modelo educativo del que -en su gran mayoría- provienen. Por tanto, una de las primeras tareas es avanzar en el significado y función de conceptos como: competencias, RA, propuestas de trabajo, criterios de evaluación, evidencias... que a la vez que les serán útiles en su desarrollo personal y profesional, son elementos clave en los procesos de evaluación a desarrollar a lo largo del curso.

Por otra parte, cabe destacar la importancia de contar con apoyos tecnológicos que faciliten el trabajo y la reflexión de alumnos y docentes. Como puede ser la herramienta SET (*Skills Evolution Tool*) creada por Tkніка (2015) con la intención de medir la evolución del grado de adquisición de las competencias del alumnado.

En síntesis, subrayamos la importancia de dar tiempo a estos procesos, de consolidar los equipos docentes -a la par que se ofrece formación continua a los mismos en relación con las temáticas que nos ocupan-, y todo ello requiere de un liderazgo claro, comprometido e inclusivo en el que se toma en cuenta y se valora la opinión y participación de todos los implicados en el proceso educativo.

## 7. Conclusiones

Las siguientes conclusiones sintetizan y ponen en relieve la importancia de los RA como elemento central para llevar a cabo una evaluación auténtica del desarrollo de competencias en titulaciones de grado, a la vez que proporcionan algunas claves para transferir la experiencia realizada a otros títulos de grado o máster.

Un cambio profundo de la evaluación de competencias, de modo que esta se realice desde los RA y no desde los instrumentos de evaluación, exige de una sensibilidad, compromiso y acción colectiva de todos los agentes implicados y, especialmente, del equipo directivo y del profesorado. Este componente actitudinal requiere, a la vez, de una conceptualización previa profunda (componente conceptual) que permita comprender, definir y asumir personal y colectivamente un enfoque propio -conjugando la evaluación del y para el aprendizaje y la evaluación como aprendizaje y evolución- que derive en un modelo y sistema de evaluación consistente y viable en una formación de calidad.

Un último componente es el procesual, desde el cuál debe organizarse y desplegar la implementación del cambio evaluativo. La complejidad propia de este tercer componente debe contemplar acciones eficientes en la aplicación exitosa del modelo y sistema definido. En este sentido la evaluación debe resituarse en el diseño y desarrollo curricular del título, de modo que los RA no sólo sean una concreción finalista de los logros en la adquisición de las competencias, sino que se adopten como referente en el diseño de las actividades de aprendizaje y estén presentes durante su realización dentro y/o fuera del aula. Para ello es imprescindible:

- a) Considerar su escalabilidad, que va desde los RA del grado a los niveles de logro en una actividad o propuesta de trabajo específica pasando por la concreción de los RA para cada curso y módulo o asignatura; y
- b) Contemplar una evaluación formativa permanente desplegada por los docentes y que, apoyada en la figura del tutor, proporcione retornos cualitativos mediante un feedback basado en las interacciones directas con los estudiantes de modo que se faciliten así mayores niveles de logro en los RA.

Esta redefinición del sistema de evaluación llevado a cabo en los Grados de Educación Infantil y Educación Primaria de HUHEZI no sólo es acorde con el contenido del protocolo de evaluación para la verificación de títulos oficiales en cuanto define y despliega “... un procedimiento general para la valoración de los resultados del aprendizaje de los estudiantes” (REACU, 2011, p. 9), sino que además es una propuesta diferenciada en el contexto universitario español de evaluación de la adquisición de competencias profesionales, en cuanto toda ella está articulada alrededor de los RA.

## Referencias

- ANECA [Agencia Nacional de Evaluación de la Calidad y Acreditación]. (2013). *Guía de apoyo para la redacción, puesta en práctica y evaluación de los resultados del aprendizaje*. ANECA.
- Astigarraga, E. y Carrera, X. (2018). Necesidades a futuro y situación actual de las competencias en Educación Superior en el contexto de España. *Revista Digital de Investigación en Docencia Universitaria*, 12(2), 35-58. <http://doi.org/10.19083/ridu.2018.731>
- Biggs, J. (2014). Constructive alignment in university teaching. *HERDSA Review of Higher Education*, 1, 5-22.
- Biggs, J. B. y Tang, C. (2011). *Teaching for quality learning at university*. McGraw-Hill Education & Open University Press.
- Brown, S. (2015). La evaluación auténtica: el uso de la evaluación para ayudar a los estudiantes a aprender. *RELIEVE*, 21(2), art M4. Recuperado de <https://ojs.uv.es/index.php/RELIEVE/article/view/7674>
- CEDEFOP [Centro Europeo de Formación Profesional]. (2017). *Defining, writing and applying learning outcomes. A European handbook*. Recuperado de [https://www.cedefop.europa.eu/files/4156\\_en.pdf](https://www.cedefop.europa.eu/files/4156_en.pdf)
- CEDEFOP [Centro Europeo de Formación Profesional]. (2019). *Los marcos de cualificaciones en Europa - Evolución en 2018, Promover la confianza mutua a partir de los resultados del aprendizaje*. Recuperado de [https://www.cedefop.europa.eu/files/9139\\_es.pdf](https://www.cedefop.europa.eu/files/9139_es.pdf)
- Comisión Europea (2009). *El Marco Europeo de Cualificaciones para el aprendizaje permanente (EQF-MEC)*. Oficina de Publicaciones de la Unión Europea.

- Frey, B. B., Schmitt, V. L. y Allen, J. P. (2012). Defining authentic classroom assessment. *Practical Assessment, Research & Evaluation*, 17(2), 1-18.
- Real Decreto 1027/2011, de 15 de julio, por el que se establece el Marco Español de Cualificaciones para la Educación Superior. BOE, texto consolidado, última modificación 7 de febrero de 2015. Recuperado de <https://www.boe.es/buscar/pdf/2011/BOE-A-2011-13317-consolidado.pdf>
- Real Decreto 96/2014, de 14 de febrero, por el que se modifican los Reales Decretos 1027/2011, de 15 de julio, por el que se establece el Marco Español de Cualificaciones para la Educación Superior (MECES), y 1393/2007, de 29 de octubre, por el que se establece la ordenación de las enseñanzas universitarias oficiales. Recuperado de <https://www.boe.es/boe/dias/2014/03/05/pdfs/BOE-A-2014-2359.pdf>
- Guerriero, S. (2017). *Pedagogical knowledge and the changing nature of the teaching*. OCDE.
- Hattie, J. (2017). "Aprendizaje visible" para profesores. *Maximizando el impacto en el aprendizaje*. Paraninfo.
- Hayward, L. (2015). Assessment is learning: the preposition vanishes. *Assessment in Education: Principles, Policy & Practice*, 22(1), 27-43. <https://doi.org/10.1080/0969594X.2014.984656>
- Hill, P. y Barber, M. (2014). *Preparing for a renaissance in assessment*. Pearsons.
- Lee, A. M. (2013). *Assessment OF... AS... FOR... learning*. Recuperado de <https://annemichellelee88.wordpress.com/2013/02/14/assessment-of-as-for-learning/>
- León, C. (2014). *Introducción a la educación superior basada en competencias*. Limusa.
- Ozaeta, A., Mongelos, A., Astigarraga, E. y Garro, E. (2018). Innovando en la universidad. Algunas claves en un proceso de cambio curricular y metodológico. En A. Villa (Ed.), *Tendencias actuales de las transformaciones de las universidades en una nueva sociedad digital* (pp. 155-166). Foro Internacional de Innovación Universitaria.
- Paniagua, A. e Istance, D. (2018). *teachers as designers of learning environments: The importance of innovative pedagogies*. OCDE.
- Pellegrino, J. W. (2017). Teaching, learning and assessing 21st century skills. En S. Guerriero (Ed.), *Pedagogical Knowledge and the changing nature of the teaching* (pp. 223-251). OCDE.
- REACU (2011). *Evaluación para la Verificación*. Protocolo de evaluación para la verificación de títulos universitarios oficiales (Grado y Máster). Recuperado de <http://www.aneca.es/Programas-de-evaluacion/Evaluacion-de-titulos/VERIFICA/Verificacion-de-Grado-y-Master/Documentacion-y-herramientas>
- Rundle, N. y Gurney, B. (2017). *Constructive alignment diagram*. Recuperado de <https://elibrary.utas.edu.au/lor/items/27385ef0-d14e-44e7-a5ca-fa6f8e501281/1/>
- Sarobe, A., López-Salas, N. y Astigarraga, E. (2019). *Evaluación del modelo de innovación ETHAZI en los centros de formación profesional de la Comunidad Autónoma del País Vasco*. Documento Interno. HUEZI.
- SantaCruz, S. (2019). *Constructively aligned assessment for deeper learning in Higher Education*. Documento Interno – Memoria de Investigación del Personal Investigador en Formación (PIF). HUEZI.
- Scott, C. L. (2015a). *El futuro del aprendizaje 2 ¿Qué tipo de aprendizaje se necesita en el siglo XXI?* Documentos de Trabajo ERF, nº 14. UNESCO.
- Scott, C. L. (2015b). *El futuro del aprendizaje 3 ¿Qué tipo de pedagogías se necesitan en el siglo XXI?* Documentos de Trabajo ERF, nº. 15. UNESCO.

Siarova, H., Sternadel, D. y Mašidlauskaitė, R. (2017). *Assessment practices for 21st century learning: review of evidence*. Oficina de Publicaciones de la Unión Europea.

Tknika. (2015). *SET (Skills Evolution Tool) Evaluación por competencias orientada hacia la evolución*. Recuperado de <https://tknika.eus/cont/proyectos/set-skills-evolution-tool/#>

Tknika. (2019). *Evaluación, feedback y calificación en el trabajo por retos ETHAZI*. Recuperado de [https://drive.google.com/file/d/1TsN8ls-q6ZESN8wXJ5p92m7nDz-4TZb\\_/view](https://drive.google.com/file/d/1TsN8ls-q6ZESN8wXJ5p92m7nDz-4TZb_/view)

## Breve Cv de los autores

### Eugenio Astigarraga Echeverría

Profesor de Educación General Básica. Licenciado en Filosofía y Ciencias de la Educación. Director Pedagógico en la empresa ALECOP, S. Coop. e integrante de la Unidad de Proyectos Educativos Internacionales. Profesor de Grado y Master de la Facultad de Humanidades y Ciencias de la Educación - Mondragon Unibertsitatea. Colaborador de Tknika - Centro de Investigación e Innovación en Formación Profesional. Doctor por Mondragon Unibertsitatea en el Programa en Innovación e Intervención Educativa (Mención Doctorado Internacional). ORCID ID: <https://orcid.org/0000-0002-1153-6420>. Email: [eastigarraga@mondragon.edu](mailto:eastigarraga@mondragon.edu)

### Arantza Mongelos García

Doctora en Psicodidáctica por la universidad del País Vasco (UPV-EHU). MSc in Applied Linguistics por la Edinburgh University. Licenciatura en Filología Inglesa (EHU-UPV) Profesora de Grado de la Facultad de Humanidades y Ciencias de la Educación - Mondragon Unibertsitatea Coordinadora de Grado de la Facultad de Humanidades y Ciencias de la Educación - Mondragon Unibertsitatea Coordinadora universitaria de proyectos internacionales en MEi (Mondragon Educación Internacional). ORCID ID: <https://orcid.org/0000-0001-5795-9173>. Email: [amongelos@mondragon.edu](mailto:amongelos@mondragon.edu)

### Xavier Carrera Farran

Doctor en Psicopedagogía. Premio extraordinario de Doctorado. Coordinador en la Universidad de Lleida del Programa de Doctorado Interuniversitario en Tecnología Educativa. Coordinador del grupo de investigación COMPETECS (Competencias, Tecnología, Educación y Sociedad) de la Universidad de Lleida [2017 SGR 1700]. Sus líneas de investigación se centran en la Tecnología Educativa; la aplicación de metodologías activas en educación y en la formación basada en competencias. ORCID ID: <https://orcid.org/0000-0003-3420-4215> Email: [carrera@pip.udl.cat](mailto:carrera@pip.udl.cat)

## Diferencias de Género y Estudios de Acceso en las Creencias del Alumnado de Grado en Educación Infantil sobre el Desarrollo de la Autonomía en el Ciclo 0-3

### Differences in Childhood Education Degree Students' Beliefs about Child Autonomy in the First Cycle (0-3 years) according to Gender and Studies of Origin

Elena Herrán Izagirre  
Nuria Galende Pérez \*  
Gorka Etxebarria Elordui

Universidad del País Vasco / Euskal Herriko Unibertsitatea, España

Este estudio tiene como objetivo profundizar en las creencias y actitudes del alumnado del Grado de Educación Infantil sobre el desarrollo de la autonomía del bebé y niño pequeño (0-3 años). La autonomía temprana se entiende como capacidad preprogramada que impulsa a la cría humana a interactuar con el mundo circundante, tanto físico como humano, de manera progresivamente más activa e independiente. En concreto, el estudio trata de analizar las diferencias en dichas creencias según el género y los estudios de procedencia, en 165 estudiantes universitarios del mencionado Grado. Las respuestas dadas a un cuestionario desarrollado al efecto se mostraron inconsistentes, por lo que se procedió a centrar las puntuaciones. Los resultados muestran diferencias por género coherentes con el carácter feminizado de la profesión y de los estudios. Asimismo, el alumnado proveniente del Grado Superior en Educación Infantil de Formación Profesional tiende a valorar mejor los aspectos prácticos de la profesión que los teóricos o de fundamentación. Esto confirma que los imaginarios de los colectivos que forman la muestra son diferentes según las dos variables elegidas, y plantea un gran reto a la formación de Grado.

**Palabras clave:** Creencia; Autonomía; Título universitario; Educación preescolar; Diferencia de sexo; Enseñanza y formación.

This study aims to go in depth in the beliefs and attitudes of the students of the Degree in Early Childhood Education in relation to the development of the autonomy of babies and toddlers (0-3 years). Early autonomy is understood as a preprogrammed capacity that encourages human breeding to interact with the surrounding world, physical and human, in a progressively more active and independent manner. Specifically, it tries to analyze the differences in these beliefs according to gender and studies of origin in 165 university students of the aforementioned Degree. The answers given to a questionnaire developed for this purpose were inconsistent, so the scores were centered. The results obtained show significant differences by gender coherent with the feminized character of the profession and of the studies. Likewise, students that come from the Higher Degree in Pre-School Education tend to value differentially the practical aspects of the profession more than the theoretical ones. This confirms that the imaginaries of the collectives that form the sample are different according to the two variables chosen, and poses a great challenge to the training in the Degree.

**Keywords:** Belief; Autonomy; Academic degree; Pre-school education; Sex difference; Teaching and training

---

\*Contacto: [nuria.galende@ehu.eus](mailto:nuria.galende@ehu.eus)

## **1. Introducción**

El ámbito científico de esta área de investigación lleva largo tiempo demostrando la importancia vital de la primera infancia en la construcción sana del psiquismo de las personas (Sánchez-Rodríguez, 2014), así como la prevalencia de dicha construcción a lo largo de la vida, condicionando, entre otras cosas, las futuras interacciones sociales (Raby et al., 2015) así como el conocimiento (Gopnik, 2010). No obstante, a pesar del inmenso volumen de datos que certifica la importancia de invertir en cuidados de calidad en este primer periodo de la vida, la evidencia científica no ha logrado calar en la sociedad, ni, lo que resulta más alarmante, en el sistema educativo en general (Herrán, Orejudo, Martínez de Morentin y Ordeñana, 2014). Esto resulta especialmente preocupante en los niveles educativos superiores (universidad, ciclos formativos), ya que son los estudios mediante los cuales se forma a los y las futuras profesionales que atenderán a bebés y niños pequeños. Los planes de estudios actuales tienden a centrarse en las capacidades cognitivas de la infancia, aplicando objetivos, contenidos, etc., propios de edades posteriores: el segundo ciclo de Educación Infantil o Primaria. Sin embargo, para que esas capacidades puedan desarrollarse es imprescindible cubrir y responder, previamente, a la necesidad humana más primordial -la construcción del psiquismo-, que, durante toda la vida, pero muy especialmente en la infancia, depende y, de hecho, se construye a partir del progresivo dominio y control del propio cuerpo.

Es lo que se conoce como autonomía, definida como la capacidad de una persona de gobernarse a sí misma, de asumir la plena responsabilidad de sus actos. Familias y profesionales de la primera infancia desean ver que los niños se hacen autónomos según esta definición, pero el camino por el que podrán llegar a serlo no siempre está claro, lo que conduce a puntos de vista muy diversos (Falk, 2018a), y más aun teniendo en cuenta el papel que ejercen las creencias de las personas adultas que trabajarán con estos niños, y que no siempre tienen una base científica.

Esa es la razón por la que conocer las creencias del alumnado del Grado de Educación Infantil y del Ciclo Superior de Educación Infantil, en tanto que guías de su futuro comportamiento profesional, es fundamental. Las tendencias de comportamiento educativo se mantienen prácticamente inalterables a pesar de los años de formación -universitaria o secundaria- debido al filtro (Lortie, 2002) que incorpora los nuevos conocimientos a los anteriores más arraigados y resistentes, deformándolos y eliminando de ellos las cuestiones disonantes, y las contradicciones paradigmáticas y conceptuales de la propia formación. Es decir, la incorporación de nuevos aprendizajes es limitada.

Todo ello subraya la responsabilidad formativa de la Universidad; es decir, la nuestra. Los cambios son complejos, pero resultan imprescindibles en la CAPV, puntera en el Estado, donde existe una tasa de escolarización temprana que se mantiene alrededor del 52% desde 2010 -según el EUSTAT (2018) la tasa de matriculación del curso 2016-17 fue del 17,5 % y 45,7 % sobre el total de 0 y 1 años, respectivamente, y del 93,5 % en las aulas de 2 años-. Esta investigación pretende dar con las claves de los ajustes imprescindibles para una Educación Infantil 0-3 años de excelencia.

## **2. Fundamentación teórica**

La autonomía temprana (Falk, 2018a; Kamii, 1982) es un ámbito central de Educación Infantil (E. I.). Es una capacidad preprogramada que impulsa a la cría humana a interactuar con el mundo circundante, físico y humano (Tomasello, 2007; Wallon, 1985), de manera progresivamente más activa e independiente, condicionando por ello los demás ámbitos de desarrollo temprano. Si no hay afecto y respeto en el trato entre el adulto y el niño, puede degenerar fácilmente en heteronomía (Kamii, 1982), "...ser gobernado por algún otro" (p. 4), o en "...las tres grandes trampas de la falsa autonomía: el condicionamiento, una exigencia de precocidad, de la que puede desprenderse una actitud de indiferencia y de abandono, y una actitud de *laissez faire*" (Falk, 2018a, p. 111). En la primera infancia la calidad de la crianza es clave, ya que apoyar la autonomía ayuda al bebé y niño pequeño a desarrollar estrategias de autocontrol (Tarullo, Obradovic y Gunnar, 2009), además de ser el mejor predictor del desarrollo de las funciones ejecutivas tempranas implicadas en él (Field, 2010; Poulton, Moffitt y Silva, 2015). El autocontrol (Moffitt et al, 2011) se asocia al éxito en la vida adulta (Field, 2010; Poulton et al, 2015) y a diferencia de la inteligencia o el estatus económico, es una variable predictora del mismo (Duckworth, 2011) y además de fácil mejora mediante intervención (Moffitt et. al., 2011).

En relación con la autonomía en E. I., se ha podido establecer cierta diferencia entre dos tipos de creencias: por una parte, creencias profesionales más tradicionales o adultocéntricas, con cierta visión deficitaria o de percepción de limitaciones en lugar de centrarse y desarrollar las potencialidades y capacidades de la primera infancia (Davis y Degotardi, 2015); y, por otra parte, tendencias más progresistas o paidocéntricas, que proponen lo que es mejor y más beneficioso, por adecuado, en cada periodo particular del desarrollo y para cada niño individual (Tonyan, Mamikonian-Zarpas y Chien, 2013). Este último tipo de creencias se ha relacionado con un cuidado más positivo y con la provisión de un entorno más rico y estimulante para los niños, como el de la educación Pikler-Lóczy (David y Appell, 2010; González-Mena, 2004; Herrán, 2013; Herrán et al, 2014), con independencia de otro tipo de variables como el tamaño de grupo (Clarke-Stewart, Vandell, Burchinal, O'Brien y McCartney, 2002). Las creencias más tradicionales parecen ser la razón por la que consolidados resultados de la investigación científica sobre la autonomía temprana no influyen en el conocimiento general ni en la práctica educativa (Pikler, 2018).

Las aproximaciones sobre la enseñanza de los docentes también están relacionadas con sus concepciones de la misma (Trigwell y Prosser, 1996); nociones, ideas previas, representaciones y creencias condicionan su práctica educativa. Así, las diferentes prácticas evaluativas universitarias evidencian diferentes prácticas docentes (Samuelowicz y Bain, 2002) que incluso pueden ser contradictorias (Ribeiro y Flores, 2016), porque los diferentes conceptos de enseñanza, aprendizaje y evaluación asociados influyen diferencialmente en cómo enseña el profesorado y en cómo aprende el alumnado (Brown, 2004; Thompson, 1992). En este sentido, y en el ámbito de las ciencias, tecnología y matemáticas, se ha incrementado la investigación sobre motivaciones, actitudes, expectativas y especialmente, concepciones, de futuros docentes (Brown, Lake y Matters, 2011; Meirik, Meijer, Verloop y Bergen, 2009), con el objetivo de diseñar propuestas metodológicas que favorezcan el cambio de sus ideas previas (Pontes, Poyato y Oliva, 2016). En relación con las representaciones y creencias del alumnado de magisterio, el

comportamiento docente se apoya en una multiplicidad de factores relacionados con una ideología o percepción de la realidad que asume el-la educador-a, y que se concretan en las múltiples dimensiones de un gran impacto en los estilos de enseñanza, en el ejercicio de la profesión, en la configuración pedagógica del quehacer educativo y en las relaciones interpersonales que se establecen dentro del aula (González-Peiteado y Pino-Yuste, 2014).

Asociada a todo lo anterior, en la formación de los futuros profesionales habría otra cuestión fundamental a tener en cuenta: el filtro intuitivo (Goodman, 1988) con el que se procesan las subsiguientes experiencias educativas y docentes (Lortie, 2002). Este filtro se constituye a partir de las creencias o representaciones sociales de cada cual, en tanto que guías básicas de pensamiento o tamiz que filtrará o no la nueva formación académica. De hecho, Akin (2013) afirma que los alumnos en formación en E. I. utilizan sus creencias propias y arraigadas para evaluar las nuevas ideas, desestimando las que chocan con ellas, tachándolas de teóricas, irrealizables o simplemente incorrectas. Entre ambas -creencias y nuevas informaciones- es imprescindible cierta congruencia (Opfer y Pedder, 2011; Opfer, Pedder, y Lavicza, 2011). De lo contrario, queda muy limitada la adquisición (Goodman, 1988; Raths, 2001) así como el provecho de las subsiguientes experiencias educativas y docentes (Lortie, 2002).

Debido a que estas creencias tienen necesariamente un origen previo al universitario y un arraigo fuerte y profundo, es oportuno hacer referencia a los estilos educativos parentales como modelo explicativo del desarrollo de las creencias sobre el despliegue de la autonomía del niño. Los estilos parentales los conforman metas de desarrollo y estrategias de socialización (Goodnow, 1985) entendidas como tendencias de comportamiento paterno, globales, estables y abiertas, con una gran repercusión y consecuencias evolutivas no solo en la infancia sino en la adolescencia. Se trata de conductas parentales relativas a la disciplina, control, dominio, apoyo, etc. (Baumrind, 1991; Erikson, 1963; Gadeyne, Ghesquière y Onghena, 2004; Rollins y Thomas, 1979). La combinación de estos aspectos ha conformado los denominados estilos educativos parentales: autoritario-recíproco, autoritario-represivo, permisivo-indulgente y permisivo negligente (MacCoby y Martin, 1983). Hoffman (1970) incluye los modelos autoritario, la retirada de afecto, y la inducción, mientras que Kellerhalls y Montandon (1997) proponen los estilos familiares contractualista, estatutario y el maternalista. Todos ellos incluyen, en diferente intensidad, reciprocidad e implicación afectiva o ausencia de ambas; control fuerte o laxo; retirada de afecto o enfado; amenaza o utilización del castigo; imposición o bien dejación o uso de la inducción y oscilan en direcciones opuestas consolidando uno u otro modelo.

Es complejo conocer la influencia de la formación académica universitaria en relación con el ámbito de conocimiento de la autonomía infantil temprana. Entre otras razones, una significativa es la existencia de contradicciones en el propio currículum del Grado de E. I., asociadas al habitual trasvase de contenidos de ciclos ulteriores a edades tempranas (Loizou y Recchia, 2018). En cualquier caso, el profesor en formación construye su propio perfil de educador informado en función del efecto transformador de las experiencias formativas del Grado en E. I., de cómo cuestionan, replantean y dan alternativa a preconcepciones y creencias anteriores sobre situaciones educativas en las que la autonomía infantil está presente y conforman la tarea del profesional de la educación temprana. Para comprender estos procesos, se elaboró una escala que pretende evaluar las pautas socializadoras y educativas concretas relacionadas con la autonomía y contextualizadas en el primer ciclo de E. I. (0-3 años). Se trata de descubrir si el

conocimiento adquirido sobre la autonomía infantil temprana y sus condiciones se refleja en las respuestas, en relación a dos variables intervinientes, asociándose a creencias informadas y complejas en lugar de a creencias más tradicionales o simplistas, propias de sus contextos previos y no formativos.

### **3. Método**

#### ***3.1. Objetivos***

Este estudio, de tipo comparativo o diferencial, pretende analizar las creencias que conforman los diferentes imaginarios de los colectivos que forman la muestra según dos variables: género y estudios de acceso al Grado en E. I. En concreto, se espera que, al tratarse de una profesión feminizada, haya diferencias en la valoración de las cuestiones asociadas al género. También se espera que el alumnado del Grado en E. I. que proviene del Grado Superior en E. I. de Formación Profesional evidencie su propio filtro, más práctico que el de la formación previa de Bachillerato.

#### ***3.2. Participantes***

Son 165 participantes, estudiantes universitarios de Grado en E. I., 86% mujeres, con edades comprendidas entre los 19 y los 66 años. El 54% de la muestra tiene la edad habitual para los estudios que cursan, 19-20 años; el 33% tiene entre 21 y 24 años y el 13% restante tiene edades superiores a los 24 años. La edad está asociada a la vía de acceso a los estudios de grado: Bachillerato, Formación Profesional y Acceso a la Universidad para mayores de 25 años. Los datos se han recogido en dos centros distintos, uno de titularidad pública y el otro privada, de la Comunidad Autónoma del País Vasco. Cabe señalar que 41 de los 165 participantes tienen cursado el Grado Superior en E. I. de Formación Profesional. El muestreo ha sido de conveniencia, seleccionando los centros y grupos/aula que ofrecían mejores garantías y condiciones de aplicabilidad, primando la validez ecológica a la representatividad estrictamente aleatoria. Como resultado de estos criterios de selección, el tamaño de la muestra final se considera suficiente para garantizar un buen nivel de “poder” en los contrastes estadísticos aplicados. Las pruebas de significación estadística se acompañan de los coeficientes de tamaño del efecto con el fin de que pueda apreciarse también la cuantía absoluta de las diferencias.

#### ***3.3. Instrumento***

Las pautas socializadoras y educativas relacionadas con la autonomía y contextualizadas en el primer ciclo (0-3 años) de E. I. a analizar se han rescatado del cuestionario CUIDANDO 0-2 (Herrán et al., 2014) aplicado a profesionales en ejercicio, adaptándolo al profesorado en formación. Así, se identifican las mismas siete dimensiones o facetas relacionadas con la autonomía infantil con sus correspondientes ítems: concepto de niño (ej.: la educación temprana es buena para la salud del niño de 0-3 años), rol de la educadora (ej.: es imprescindible la intervención del adulto para solucionar los conflictos entre niños), actividad diaria (ej.: a la hora de la siesta todos los niños deberían dormir), sentimientos asociados a la actividad (ej.: creo que mis abrazos, besos y caricias son imprescindibles para el bienestar del niño), interacción con los niños (ej.: disfrutaría del momento del cambio de pañal a los niños/as de 0-3 años.), relación con la pareja educativa y evaluación (ej.: es necesario negociar y compartir los objetivos del primer ciclo (0-3 años) con la pareja

pedagógica), evaluación, innovación y mejora (ej.: compartiré a diario las observaciones con mis compañeras).

Así, se han diseñado 35 ítems con los que se pretende indagar sobre las creencias relativas a las pautas educativas y de crianza autónomas del alumnado del grado en tanto que futuros profesionales de la primera infancia. Todos los ítems se ordenarían según las dos modalidades de intervención: directa e indirecta. La primera reúne exclusivamente los ítems relacionados con el hacer individual en vivo y en directo a los bebés y niños pequeños, mientras que la segunda incluiría todo lo demás: contexto educativo, opiniones, valoraciones o perfil educador. En el proceso de construcción del instrumento se contó con el criterio de 20 profesionales expertos, que aportaron validez de contenido a la elaboración y selección de los ítems.

#### **3.4. Procedimiento de recogida y análisis de datos**

En primer lugar, se realizaron los análisis descriptivos básicos, encontrando coeficientes de asimetría y curtosis que mostraron distribuciones muy alejadas de la normalidad en los ítems 7, 17, 19, 26, 31, 33, 34. Estos ítems se retiraron de los análisis posteriores para evitar sesgos importantes en la estimación de los coeficientes.

Los primeros análisis descriptivos y exploratorios mostraron con claridad que el instrumento no medía de manera consistente las dimensiones postuladas. Se exploraron diferentes elementos que explicaran estos resultados, estableciéndose como posibilidad la presencia de sesgos importantes como la aquiescencia, deseabilidad social, radicalidad en las respuestas (Morales, 2000).

Como estrategia para resolver o paliar este problema, se procedió al centrado de las respuestas, utilizando la media de cada sujeto en cada ítem como norma individual (Schwartz, 2003). Con ello, no se trabaja con las puntuaciones brutas, sino con las relativas a la media de cada uno de los sujetos. La puntuación en el ítem refleja, por tanto, si dicho ítem se encuentra entre los más o menos valorados por los participantes, en lugar de su valoración en términos absolutos.

Dado que la orientación del ítem ha mostrado un comportamiento diferencial en numerosos estudios previos (Solís, 2015), se analizan por separado los ítems orientados positivamente o favorables a la autonomía temprana y los orientados negativamente o favorables a la heteronomía. El primer grupo lo conforman 15 ítems -6 directos y 9 indirectos- relativos a contenidos trabajados en el Grado en E. I., o creencias informadas sobre la autonomía temprana, fruto de la investigación en primera infancia. El segundo, formado por 13 ítems -6 directos y 7 indirectos- agrupa contenidos y creencias ajenas a los contenidos del Grado y por ello, creencias más simplistas que trae el alumnado, asociadas a la heteronomía temprana.

Para constatar posibles diferencias según las variables criterio mencionadas, se utilizó el Análisis de la Varianza, que identificó la existencia o no de diferencias estadísticamente significativas entre los grupos. Para estimar el tamaño del efecto, se utilizó el coeficiente *eta*. Los cuadros se presentan ordenadas por tamaño del efecto de manera que se puede identificar con claridad en qué ítems aparecen diferencias de mayor tamaño.

## 4. Resultados y conclusiones

El cuadro 1 presenta las diferencias encontradas según el género en los ítems orientados hacia el desarrollo de la autonomía.

Cuadro 1. Puntuaciones centradas de los ítems orientados hacia la autonomía según el género

	MEDIA			F	SIG.	ETA <sup>2</sup>
	MUJER	HOMBRE	TOTAL			
24. Dejaría a mis hijos/as en manos de mis propios compañeros/as de Grado /Ciclo.	-,58	,14	-,48	8,297	,005	,050
06. En el período de adaptación haré todo lo que esté en mis manos para que los niños/as no lloren cuando se separen de sus padres.	,04	-,59	-,04	5,993	,015	,036
25. <i>Para ser un buen profesional es más importante ser reflexivo que práctico.*</i>	-1,02	-,50	-,95	5,621	,019	,034
28. Es importante planificar el día a día de la escuela en torno a los cuidados (cambio de pañal, higiene, alimentación...).	,81	,50	,77	3,619	,059	,022
22. Compartiré a diario mis observaciones con mis compañeras/os.	,13	,36	,16	1,405	,238	,009
18. Me gustará tratar con los/las padres/familias de los niños y niñas.	,67	,50	,65	1,075	,301	,007
32. En el aula de dos años la tutora debe cambiar pañales al igual que la auxiliar.	,49	,68	,52	1,013	,316	,006
11. Disfrutaría del momento del cambio de pañal a los niños/as de 0-3 años.	-,35	-,59	-,39	1,005	,318	,006
20. Me preocupa si sabré gestionar adecuadamente en el día a día la acogida y despedida de los niños.	-,57	-,77	-,60	,623	,431	,004
35. Es necesario negociar y compartir los objetivos del primer ciclo (0-3 años) con la pareja pedagógica.	,83	,91	,84	,385	,536	,002
21. Soy capaz de realizar mi trabajo en el ciclo 0-3 años.	,34	,42	,35	,196	,659	,001
13. Antes de quitarle los mocos a un niño/a le avisaría de lo que le voy a hacer.	,73	,66	,72	,192	,662	,001
10. En las comidas dejaría que los propios niños eligiesen la cantidad que van a comer.	-1,39	-1,50	-1,40	,173	,678	,001
04. Dejar al niño que adopte la postura que quiera en el cambio de pañal refuerza su seguridad.	-,64	-,68	-,65	,019	,891	,000
29. Cuando un niño que quita a otro un juguete tenemos que recordarle que no lo puede hacer mientras esté jugando con él.	,51	,50	,51		,969	,000

Mujeres (n=139) Hombres (n=22) Total (n=161)

\* Los ítems indirectos van en cursiva en todas los cuadros para facilitar la lectura.

Fuente: Elaboración propia.

Encontramos tres ítems con diferencias estadísticamente significativas ( $p < ,05$ ), dos directos -24, 06- y uno indirecto -25-. En el ítem 24 -dejar a los propios hijos con los colegas- las mujeres otorgan una valoración muy negativa (-,58) mientras que los hombres otorgan una valoración media (,14). La valoración del ítem 06 -preocupación por el periodo de adaptación- es media en el caso de las mujeres (,04) y negativa en el caso de los hombres (-,59). En el ítem indirecto -25-, la valoración sobre ser reflexivo frente a práctico es negativa en ambos colectivos, pero extremadamente negativa entre las mujeres (-1,02) frente a los hombres (-,50).

Los resultados señalarían, pues, algunas diferencias asociadas al género. Así, las mujeres mostrarían una evidente desconfianza hacia sus compañeros, lo que podría ser coherente con el carácter marcadamente feminizado de la profesión (Anliak y Beyazkurk, 2008), además de relacionarse con el estilo millennial (Clark y Byrnes, 2015), que se siente por

encima de la media en diferentes aspectos. También éstas se preocuparían poco en la adaptación, aunque bastante más que los hombres, y se decantarían diferencialmente por la practicidad frente a la reflexión (Avgitidou, Pnevmatikos, y Likomitrou, 2013; Vartuli y Rohs, 2009).

El cuadro 2 presenta las diferencias encontradas según el género en los ítems orientados hacia el desarrollo de la heteronomía.

Cuadro 2. Puntuaciones centradas de los ítems orientados hacia la heteronomía según género

	MEDIA			F	SIG.	ETA <sup>2</sup>
	MUJER	HOMBRE	TOTAL			
02. Me tranquilizaría establecer normas estrictas en el aula para evitar el caos.	-,97	-,45	-,90	5,139	,025	,031
12. A la hora de comer me quedaría realmente tranquila/o cuando viese que el/la niño/a ha terminado lo que tenía en el plato.	,45	,00	,39	4,656	,032	,028
14. En mi opinión, para que un niño/a aprenda a controlar los esfínteres, es adecuado ponerle a la misma hora en el orinal	,20	-,27	,13	4,583	,034	,028
30. Aunque la adaptación sea mala el niño pueden construir un buen vínculo sano con la educadora	,61	,14	,54	3,601	,060	,022
08. Si sientan al niño/a antes de estar preparado, yo no lo sentiría en el aula.	-,22	,17	-,17	1,800	,182	,011
01. Es básico que el niño/a se relacione, desde bebé, con la mayor cantidad de educadores/as posible desde el inicio de su escolarización.	-1,20	-,93	-1,17	1,165	,282	,007
23. En el prácticum se aprende más sobre la educación de los niños y las niñas de esta edad que en todo el resto del grado/ciclo.	,74	,59	,72	,437	,510	,003
03. Es más importante respetar el movimiento libre del niño en los momentos de actividad autónoma que durante los cuidados (cambio de pañal, higiene, alimentación...).	,03	,18	,05	,278	,599	,002
15. A la hora de la siesta todos los niños/as deberían dormir.	-,63	-,54	-,61	,102	,750	,001
27. Debemos mostrar nuestro enfado cuando los niños/as actúan mal.	-,02	,05	-,01	,085	,771	,001
09. Es necesaria la intervención de la educadora para resolver los conflictos infantiles.	,08	,13	,09	,055	,815	,000
05. La escolarización temprana (0-3) es buena para la salud de los niños/as.	,06	,04	,06	,005	,944	,000
16. Considero que mis abrazos, besos y caricias son imprescindibles para el bienestar del niño.	,87	,86	,87	,003	,959	,000

\* Mujeres (n=139) Hombres (n=22) Total (n=161).

Fuente: Elaboración propia.

Entre los ítems orientados hacia la heteronomía, en relación a la variable género (cuadro 2), encontramos tres ítems directos -02, 12 y 14-, con diferencias estadísticamente significativas ( $p < 0,05$ ). La valoración del ítem 02 -normas estrictas para evitar el caos- es extremadamente negativa entre las mujeres (-,97) mientras que entre los hombres (-,45) es bastante negativa. En el ítem 12 las mujeres valoran positivamente que el niño acabe el plato (,45) mientras que los hombres tienen una posición media al respecto (,00). En el ítem 14 se mantiene esta tendencia: las mujeres valoran de manera moderadamente positiva que se coloque al niño a la misma hora en el orinal para aprender a controlar esfínteres (,20), mientras que los hombres valoran de manera moderadamente negativa esta cuestión (-,27).

Las valoraciones de estos tres ítems tienen que ver con la norma y el aprendizaje autoritario de hábitos. Las mujeres serían más permisivas que los hombres con el caos del aula, pero más estrictas en lo relativo a terminar lo servido en el plato o el condicionamiento para el control de esfínteres. Evidentemente, son cuestiones relacionadas con el género. Aunque todo apunta a que, al ser tareas normalmente desarrolladas por la madre en el entorno familiar, habría que indagar si se trata de una cuestión atávica femenina asociada a la ancestral idea de comida-amor (Hamburg, Finkenauer y Schuengel, 2014) o a la exigencia de la limpieza precoz (Falk y Vincze, 2018), y aprendida de esos estilos educativos parentales por el alumnado femenino.

En resumen, los resultados asociados al género señalan algunas diferencias relevantes. Así, los ítems orientados hacia la autonomía dejan translucir que las mujeres mostrarían en conjunto cierta rigidez y exigencia autoritaria (Hoffman, 1970; MacCoby y Martín, 1983) en forma de evidente desconfianza profesional para con sus propios compañeros, escasa preocupación por la inevitable adaptación -aunque bastante más que los hombres en los que a la vista de los demás ítems este hecho podría deberse a dejadez o impotencia (Falk, 2018a)- y nuevamente de practicidad frente a reflexión; cuestiones todas ellas coherentes con el carácter marcadamente feminizado de la profesión, además de con la superioridad del estilo millennial y a la confusión entre realidad y práctica, anteriormente citados. Por su parte, las valoraciones de los tres ítems orientados a la heteronomía seguirían teniendo que ver con la norma autoritaria. Mantienen una diferencia similar por géneros, aunque el primero se decanta en dirección contraria a los otros dos. Así, las mujeres serían más permisivas que los hombres con el caos del aula -inevitable e inherente a la actividad de juego temprano- pero diferencialmente reticentes a la gestión autónoma del hambre y la saciedad (Vincze, 2018) o proclives a promover el control de esfínteres mediante condicionamiento (Falk y Vincze, 2018). Convendría indagar si se trata de una cuestión atávica femenina asociada a la ancestral idea de comida-amor (Hamburg et al. 2014) o a la exigencia de la limpieza precoz, y aprendida de los estilos educativos parentales (Hoffman, 1970; MacCoby y Martín, 1983) por el alumnado femenino, al ser tareas normalmente desarrolladas por la madre en el entorno familiar.

En cuanto a las posibles diferencias según formación previa (Bachillerato vs. FP), el cuadro 3 presenta las encontradas en los ítems orientados hacia el desarrollo de la autonomía.

Como puede observarse (cuadro 3), seis ítems presentan diferencias estadísticamente significativas ( $p < 0,05$ ) entre el alumnado procedente de Bachiller y el procedente de FP, de los que tres serían directos -11, 06, 29- y otros tres, indirectos -21, 20, 25-.

Entre los primeros, en el 11 -disfrute del cambio de pañal- el alumnado procedente de FP muestra una valoración media y el alumnado procedente de Bachillerato muestra una valoración claramente negativa. Aunque las valoraciones del cambio de pañal son bajas en ambos grupos, la diferencia entre ellas puede deberse a que se trata de un ámbito curricular de la FP y no del grado. Con el ítem 06 -sobre el proceso de adaptación- ocurre lo contrario: los alumnos de Bachillerato otorgan una puntuación media, mientras que los de FP otorgan una valoración claramente negativa (.05 frente a -.41). El ítem 29 se refiere a la gestión de las normas asociadas a la socialización de los juguetes y recibe una valoración positiva en ambos colectivos, pero significativamente superior en el alumnado de Bachiller (.61 frente a .32 del de FP). De los tres ítems indirectos, en el ítem 21, referido a la propia capacidad para realizar el trabajo, el alumnado de Bachiller otorga una valoración menor

(,25) que el de FP (,59), confirmando que estos últimos se sienten más capaces de realizar su trabajo en el ciclo 0-3 años. La valoración del ítem 20 apunta a que a ninguno de los grupos le preocupa la gestión de las transiciones diarias, pero diferencialmente menos al alumnado de FP (-,92) que a los de Bachiller (-,45). Finalmente, la valoración sobre ser reflexivo frente a práctico -ítem 25- es muy baja en ambos colectivos, pero bastante más baja entre el alumnado de FP (-1,23) frente a los de Bachiller (-,80).

Cuadro 3. Puntuaciones centradas de los ítems orientados hacia la autonomía según formación previa

	MEDIA			F	SIG.	ETA <sup>2</sup>
	BACHILLER	FP	TOTAL			
11. Disfrutaría del momento del cambio de pañal a los niños/as de 0-3 años.	-,58	,08	-,40	13,205	,000	,082
21. Soy capaz de realizar mi trabajo en el ciclo 0-3 años.	,25	,59	,34	7,608	,007	,050
20. Me preocupa si sabré gestionar adecuadamente en el día a día la acogida y despedida de los niños.	-,45	-,92	-,58	5,904	,016	,039
25. Para ser un buen profesional es más importante ser reflexivo que práctico.	-,80	-1,23	-,92	5,832	,017	,038
06. En el período de adaptación haré todo lo que esté en mis manos para que los niños/as no lloren cuando se separen de sus padres.	,05	-,41	-,08	4,881	,029	,032
29. Cuando un niño quita a otro un juguete tenemos que recordarle que no lo puede hacer mientras esté jugando con él.	,61	,32	,53	4,030	,047	,027
04. Dejar al niño que adopte la postura que quiera en el cambio de pañal refuerza su seguridad.	-,82	-,43	-,71	3,560	,061	,024
10. En las comidas dejaría que los propios niños eligiesen la cantidad que van a comer.	-1,51	-1,26	-1,44	1,547	,216	,010
28. Es importante planificar el día a día de la escuela en torno a los cuidados (cambio de pañal, higiene, alimentación...).	,81	,71	,79	,580	,447	,004
13. Antes de quitarle los mocos a un niño/a le avisaría de lo que le voy a hacer	,75	,71	,74	,137	,711	,001
24. Dejaría a mis hijos/as en manos de mis propios compañeros/as de Grado /Ciclo.	-,48	-,41	-,46	,129	,721	,001
35. Es necesario negociar y compartir los objetivos del primer ciclo (0-3 años) con la pareja pedagógica.	,82	,86	,83	,121	,728	,001
32. En el aula de dos años la tutora debe cambiar pañales al igual que la auxiliar.	,52	,57	,53	,104	,748	,001
22. Compartiré a diario mis observaciones con mis compañeras/os.	,19	,18	,19	,011	,918	,000
18. Me gustará tratar con los/las padres/familias de los niños y niñas.	,63	,64	,63	,008	,930	,000

\* Bachiller (n=108); FP (n=41); Total (n=149).

Fuente: Elaboración propia.

Estos resultados evidenciarían cierta fractura entre ambos grupos. Así, el alumnado de FP, a pesar de la formación universitaria compartida con los alumnos provenientes de Bachiller, presentaría un filtro intuitivo (Goodman, 1988) que le haría más resistente a responder a las dificultades de la adaptación o a la socialización de los juguetes, y más proclive a disfrutar más del cambio de pañal -aunque poco-, a sentirse capaces de desarrollar su tarea, a despreocuparse de las entradas y salidas y a valorar la practicidad frente a la reflexión, todo lo cual cuestionaría el objetivo formativo del propio Grado en E. I. Además, esta tendencia podría asociarse a la autosuficiencia del estilo millennial (Clark y Byrnes, 2015), y también a la habitual confusión entre realidad y práctica, que en las creencias simplistas se interpretan como sinónimos, mientras que las más progresistas comparten que un buen análisis y reflexión de la realidad, que incorpora elementos

teóricos relevantes, se considera lo único que puede guiar una práctica realmente transformadora (Avgitidou et al, 2013; Vartuli y Rohs, 2009).

Cuadro 4. Puntuaciones centradas de los ítems orientados hacia la heteronomía según formación previa

	MEDIA			F	SIG.	ETA <sup>2</sup>
	BACHILLER	FP	TOTAL			
09. Es necesaria la intervención de la educadora para resolver los conflictos infantiles.	,26	-,29	,11	9,340	,003	,060
23. En el Practicum se aprende más sobre la educación de los niños y las niñas de esta edad que en todo el resto del grado/ciclo.	,62	1,10	,75	7,259	,008	,047
01. Es básico que el niño/a se relacione, desde bebé, con la mayor cantidad de educadores/as posible desde el inicio de su escolarización.	-1,15	-1,56	-1,26	4,879	,029	,032
08. Si sientan al niño/a antes de estar preparado, yo no lo sentaría en el aula.	-,28	,08	-,18	2,408	,123	,016
14. En mi opinión, para que un niño/a aprenda a controlar los esfínteres, es adecuado ponerle a la misma hora en el orinal.	,07	,34	,15	2,388	,124	,016
12. A la hora de comer me quedaría realmente tranquila/o cuando viese que el/la niño/a ha terminado lo que tenía en el plato.	,47	,32	,43	,809	,370	,005
15. A la hora de la siesta todos los niños/as deberían dormir.	-,53	-,68	-,57	,502	,480	,003
05. La escolarización temprana (0-3) es buena para la salud de los niños/as.	,03	,10	,05	,180	,672	,001
16. Considero que mis abrazos, besos y caricias son imprescindibles para el bienestar del niño.	,85	,91	,87	,121	,728	,001
03. Es más importante respetar el movimiento libre del niño en los momentos de actividad autónoma que durante los cuidados (cambio de pañal, higiene, alimentación...).	,04	-,02	,03	,079	,780	,001
02. Me tranquilizaría establecer normas estrictas en el aula para evitar el caos.	-,90	-,85	-,88	,070	,791	,000
30. Aunque la adaptación sea mala el niño pueden construir un buen vínculo sano con la educadora.	,49	,54	,50	,060	,807	,000
27. Debemos mostrar nuestro enfado cuando los niños/as actúan mal.	-,00	,00	,00	,000	,983	,000

\* Bachiller (n=108); FP (n=41); Total (n=149).

Fuente: Elaboración propia.

Recordemos que los ítems del cuadro 4 se denominan orientados hacia la heteronomía, puesto que las conductas que reflejan serían contrarias al desarrollo de la autonomía. En cuanto a su relación con la variable formación previa, encontramos tres ítems indirectos - 09, 23, 01- con diferencias estadísticamente significativas ( $p < ,05$ ).

En el ítem 09 la valoración es media-baja para el alumnado de Bachillerato (.26), y bastante más baja para el de FP (-.29). Es decir, mientras que para el alumnado de Bachiller es adecuado que la educadora intervenga en los conflictos, para el alumnado de FP, no, lo que invierte el sentido del ítem, resultando favorable a la autonomía. El ítem 23 – aprendizaje en el prácticum vs. Grado– presenta una valoración positiva media-alta en el caso de los alumnos de Bachiller (.62) y muy alta (1,10) para el alumnado de FP, lo que supone que ambos grupos valoran más la formación del practicum que la del resto del Grado, pero esta tendencia es extremadamente alta en el caso de los estudiantes que proceden de FP. Finalmente, en el ítem 01 -que bebés y niños pequeños se relacionen con la menor cantidad de educadores posible en el tiempo de la escolarización temprana- se va a producir una inversión, ya que la valoración en ambos grupos es negativa, incluso muy

negativa para el de FP (-1,56), lo que vuelve a transformar el sentido del ítem en autónomo.

Los resultados informan de que los estudiantes con formación previa en FP invierten el sentido de dos de los ítems, transformándolos en favorables a la autonomía: tienen bastante más claro que los de Bachiller que en la mayoría de los conflictos tempranos no es necesaria la intervención adulta (Tardos y Vasseur-Paumelle, 2018). También saben mejor que las relaciones vinculares en la educación en colectividad deben ser con pocas personas adultas (Falk 2013, 2018b). Nuevamente, la practicidad se impone a la reflexión, aunque diferencialmente, ya que la valoración del alumnado de FP mucho más alta que la de Bachiller (Avgitidou et al., 2013; Vartuli y Rohs, 2009).

En definitiva, los resultados asociados a los estudios de acceso al Grado informan de que entre ambas muestras habría una fractura asociada al filtro intuitivo (Goodman, 1988), que transluce normatividad, sentido del deber o incluso cierto fatalismo (Kellerhalls y Montandon, 1997). Así, el alumnado de FP, a pesar de la formación universitaria compartida con los alumnos provenientes de Bachiller, sería más resistente a responder a las dificultades de la adaptación, a la gestión de la acogida, o a la socialización autónoma de los juguetes, cuestiones inevitables por inherentes a la escolarización temprana, y más proclives a sentirse capaces de hacer su tarea y a vindicar la practicidad frente a la reflexión, cuestionando por todo ello el objetivo formativo del propio Grado en E. I. Esta tendencia, además de asociarse a la autosuficiencia del estilo millennial (Clark y Byrnes, 2015) -se sentirían capaces de hacer algo de lo que se desentienden- podría deberse a la confusión entre realidad y práctica, que en las creencias simplistas se interpretan como sinónimos, mientras que las más progresistas comparten que un buen análisis y reflexión de la realidad, que incorpora elementos teóricos relevantes, se considera lo único que puede guiar una práctica realmente transformadora (Avgitidou et al, 2013; Vartuli y Rohs, 2009). Además, la diferencia en la valoración del cambio de pañal entre ambos colectivos -poco favorable para la FP y bastante desfavorable para Bachiller- evidencia que es un ámbito curricular en la FP y no en el Grado. Los ítems heterónomos informarían de cierta congruencia entre el filtro y la nueva información (Opfer y Pedder, 2011; Opfer et al, 2011): el alumnado proveniente de FP tendría bastante más claro que el de Bachiller que en la mayoría de los conflictos tempranos no sería necesaria la intervención adulta (Tardos y Vasseur-Paumelle, 2018) y también que las relaciones vinculares seguras y estrechas en educación temprana deberían ser con pocas personas adultas (Falk 2013, 2018b).

En síntesis, este estudio diferencial confirma que los imaginarios de los colectivos que forman la muestra son diferentes según las dos variables elegidas: género y estudios de acceso al Grado de Maestro E. I. Los ítems que presentan diferencias significativas son 15, de los que dos se repiten en dos de las variables -6 y 25- y se refieren a diferentes cuestiones normativas: al compromiso con la tarea educativa temprana y a la dicotomía teoría-práctica.

Además, el alumnado con la formación actual tendría importantes dificultades para deshacer su filtro intuitivo sobre la educación temprana, mayormente normativo y autoritario, y para tomar conciencia de la relevancia de fomentar adecuadamente la autonomía temprana en su tarea diaria, además de aprender a hacerlo en la práctica. Esto nos lleva a plantear un gran reto a la formación de Grado, ya que además de cuestionarse sus contenidos específicos y eliminar los de niveles educativos ulteriores, debe acceder a

las creencias que conforman los diferentes filtros intuitivos y deshacerlos para poder ofrecer al alumnado una formación realmente informada, veraz y actualizada.

## Referencias

- Akin, Z. (2013). Examining the beliefs of Turkish preservice Early Childhood teachers regarding early childhood curriculum. *Journal of Research in Childhood Education*, 27, 302-318. <https://doi.org/10.1080/02568543.2013.796331>
- Anliak S. y Beyazkurk D. S. (2008). Career perspectives of male students in early childhood education. *Educational Studies*, 34(4), 309-317. <https://doi.org/10.1080/03055690802034518>
- Avgitidou, S., Pnevmatikos, D. y Likomitrou, S. (2013). Preservice teachers' beliefs about childhood: challenges for a participatory early childhood education? *Journal of Early Childhood Teacher Education*, 34(4), 390-404. <https://doi.org/10.1080/10901027.2013.845633>
- Baumrind, D. (1991). Parenting styles and adolescent development. En R. M. Lerner, A. C. Petersen y J. Brooks-Gunn (Eds.), *Encyclopedia of adolescence*, vol. 2 (pp. 746-758). Nueva York, NY: Garland Publishing.
- Brown, G. (2004). Teachers' conceptions of assessment: Implications for policy and professional development. *Assessment in Education: Principles, Policy & Practice*, 11(3), 301-318. <https://doi.org/10.1080/0969594042000304609>
- Brown, G. T., Lake, R. y Matters, G. (2011). Queensland teachers' conceptions of assessment: The impact of policy priorities on teacher attitudes. *Teaching and Teacher Education: An International Journal of Research and Studies*, 27(1), 210-220. <https://doi.org/10.1016/j.tate.2010.08.003>
- Clark, S. y Byrnes, D. (2015). What Millennial Preservice Teachers Want to Learn in Their Training. *Journal of Early Childhood Teacher Education*, 36(4), 379-395. <https://doi.org/10.1080/10901027.2015.1100148>
- Clarke-Stewart, K. A., Vandell, D. L., Burchinal, M., O'Brien, M. y McCartney, K. (2002). Do regulable features of child-care homes affect children's development? *Early Childhood Research Quarterly*, 17, 52-86. [https://doi.org/10.1016/S0885-2006\(02\)00133-3](https://doi.org/10.1016/S0885-2006(02)00133-3).
- David, M. y Appell, G. (2010). *Lóczy. Una insólita atención personal*. Barcelona: Octaedro.
- Davis, B. y Degotardi, S. (2015). Educators' Understandings of, and Support for, Infant Peer Relationships in Early Childhood Settings. *Journal of Early Childhood Research*, 13(1), 64-78. <https://doi.org/10.1476718X14538600>
- Duckworth, A. L. (2011). The significance of self-control. *Proceedings of the National Academy of Sciences*, 108(7), 2639-2640. <https://doi.org/10.1073/pnas.1019725108>
- Erikson, E. H. (1963). *Infancia y Sociedad*. Ediciones Horme.
- Falk, J. (2013). Si tocamos el cuerpo del bebé. En J. Falk (Ed.), *Bañando al bebé. El arte del cuidado* (pp. 7-16). Asociación Pikler-Lóczy de Hungría.
- Falk, J. (2018a). Los fundamentos de la verdadera autonomía. En E. Herrán (Ed.), *Claves de la educación Pikler-Lóczy: Compilación de 20 artículos escritos por sus creadoras* (pp. 89-114). Asociación Pikler-Lóczy de Hungría.

- Falk, J. (2018b). Claves de la continuidad en la educación de los niños que viven en una casa cuna. Adaptación, continuidad y salida de la institución. En E. Herrán (Ed.), *Claves de la educación Pikler-Lóczy: Compilación de 20 artículos escritos por sus creadoras* (pp. 277-344). Asociación Pikler-Lóczy de Hungría.
- Falk, J. y Vincze, M. (2018). El desarrollo del control de esfínteres y el interés del niño pequeño hacia las funciones corporales. En E. Herrán (Ed.), *Claves de la educación Pikler-Lóczy: Compilación de 20 artículos escritos por sus creadoras* (pp. 163-178). Asociación Pikler-Lóczy de Hungría.
- Field, F. (2010). *The foundation years: preventing poor children becoming poor adults. The report of the independent review on poverty and life chances*. Cabinet Office.
- Gadeyne, E., Ghesquiere, P. y Onghena, P. (2004). Longitudinal relations between parenting and child adjustment in young children. *Journal of Clinical Child and Adolescent Psychology*, 33(2), 347-358. [https://doi.org/10.1207/s15374424jccp3302\\_16](https://doi.org/10.1207/s15374424jccp3302_16)
- González-Mena, J. (2004). What can an orphanage teach us? Lessons from Budapest. *Young Children*, 59(5), 26-29.
- González-Peiteado, M. y Pino-Yuste, M. (2014). Aproximación a las representaciones y creencias del alumnado de magisterio sobre los estilos de enseñanza. *Educación XXI*, 17(1), 83-110.
- Goodman, J. (1988). Constructing a practical philosophy of teaching: A study of preservice teachers' professional perspectives. *Teaching and Teacher Education*, 4(2), 121-137. [https://doi.org/10.1016/0742-051X\(88\)90013-3](https://doi.org/10.1016/0742-051X(88)90013-3)
- Goodnow, J. J. (1985). Change and variation in ideas about childhood and parenting. En I. E. Sigel, (Ed.), *Parental belief systems: The psychological consequences for children* (pp. 235-270). Erlbaum.
- Gopnik, A. (2010). How babies think. *Scientific American*, 76-81.
- Hamburg, M.E., Finkenauer, C. y Schuengel, C. (2014). Food for love: The role of food offering in empathic emotion regulation. *Frontiers in Psychology*, 5. <https://doi.org/10.3389/fpsyg.2014.00032>
- Herrán, E. (2013). La educación Pikler-Lóczy. Cuando educar empieza por cuidar. *Revista Latinoamericana de Educación Infantil*, 2(3), 37-56.
- Herrán, E., Orejudo, S., Martínez de Morentín, J. I. y Ordeñana, B. (2014). Actitudes docentes y autonomía en Educación Infantil 0-2: Un estudio exploratorio en la Comunidad Autónoma del País Vasco (CAPV). *Revista de Educación*, 365, 150-176.
- Hoffman, M. L. (1970). Conscience, personality and socialization techniques. *Human Development*, 13, 90-126. <https://doi.org/10.1159/000270884>
- Kamii, K. (1982). La autonomía como objetivo de la educación: implicaciones de la teoría de Piaget. *Infancia y Aprendizaje*, 18, 3-32. <https://doi.org/10.1080/02103702.1982.10821934>
- Kellerhalls, J. y Montandon, C. (1997). Les styles éducatifs. En F. De Singly (Dir.), *La famille l'état des savoirs* (pp. 194-200). Éditions La Découverte.
- Loizou, E. y Recchia, S. (2018). In-Service Infant Teachers Re-Envision Their Practice Through a Professional Development Program. *Early Education and Development*, 29(1), 91-103. <https://doi.org/10.1080/10409289.2017.1343561>
- Lortie, D. C. (2002). *Schoolteacher: A sociological study*. University of Chicago Press.

- Maccoby, E. E. y Martin, J. A. (1983). Socialization in the context of the family: parent-child interaction. En E. M., Hetherington & P. H., Mussen (Eds.), *Handbook of child psychology: vol. 4. Socialization, personality and social development* (pp. 1-101). Wiley.
- Meirink, J., Meijer, P., Verloop, N. y Bergen, T. (2009). Understanding teacher learning in secondary education: The relations of teacher activities to changed beliefs about teaching and learning. *Teaching and Teacher Education*, 25(1), 89-100.  
<https://doi.org/10.1016/j.tate.2008.07.003>.
- Moffitt, T. E., Arseneault, L., Belsky, D., Dickson, N., Hancox, R. J., Harrington, H. L., Houts, R., Poulton, R., Roberts, B., Ross, S., Sears, M., Thomson, M. y Caspi, A. (2011). A gradient of childhood self-control predicts health, wealth, and public safety. *PNAS*, 108, 2693-2698.  
<https://doi.org/10.1073/pnas.1010076108>
- Morales, P. (2000). *Medición de actitudes en psicología y educación*. Universidad Pontificia Comillas.
- Opfer, V. D. y Pedder, D. (2011). Conceptualizing teacher professional learning. *Review of Educational Research*, 81(3), 376-407. <https://doi.org/10.3102/00346543111413609>
- Opfer, V. D., Pedder, D. G. y Lavicza, Z. (2011). The role of teachers' orientation to learning in professional development and change: a national study of teachers in England. *Teaching and Teacher Education*, 27(2), 443-53. <https://doi.org/10.1016/j.tate.2010.09.014>
- Pikler, E. (2018). La competencia del bebé. En E. Herrán (Ed.), *Claves de la educación Pikler-Lóczy: Compilación de 20 artículos escritos por sus creadoras* (pp. 59-72). Asociación Pikler-Lóczy de Hungría.
- Pontes, A., Poyato, F. J. y Oliva, J. M. (2016). Concepciones sobre evaluación en la formación inicial del profesorado de ciencias, tecnología y matemáticas. *Revista Iberoamericana de Evaluación Educativa*, 9(1), 91-107. <https://doi.org/10.15366/riee2016.9.1.006>
- Poulton, R., Moffitt, T. E. y Silva, P. A. (2015). The Dunedin Multidisciplinary Health and Development Study: overview of the first 40 years, with an eye to the future. *Social Psychiatry and Psychiatric Epidemiology*, 50(5), 679-693.  
<https://doi.org/10.1007/s00127-015-1048-8>
- Raby, K. L., Lawler, J. M., Shlafer, R. J., Hesemeyer, P. S., Collins, W. A. y Sroufe, L. A. (2015). The interpersonal antecedents of supportive parenting: a prospective, longitudinal study from infancy to adulthood. *Developmental Psychology*, 51(1), 115-123.  
<https://doi.org/10.1037/a0038336>
- Raths, J. (2001). Teachers' beliefs and teaching beliefs. *Early Childhood Research & Practice*, 3(1). Recuperado de <https://bit.ly/2DXyLB3>
- Ribeiro, D. y Flores, M. A. (2016). Conceptions and Practices of Assessment in Higher Education: A Study of Portuguese University Teachers. *Revista Iberoamericana de Evaluación Educativa*, 9(1), 9-29. <https://doi.org/10.15366/riee2016.9.1.001>
- Rollins, B. C. y Thomas, D. L. (1979). Parental support, power and control techniques in the socialization of children. En E. R. Burr et al. (Eds.), *Contemporary theories about the family* (pp. 317-364). Free Press.
- Samuelowicz, K. y Bain, J. D. (2002) Identifying academics' orientations to assessment practice. *Higher Education*, 43(2), 173-201. <https://doi.org/10.1023/A:1013796916022>
- Sánchez-Rodríguez, J. (2014). La intervención desde la psicomotricidad relacional en la psicosis infantil. *Revista Iberoamericana de Psicomotricidad y Técnicas Corporales*, 39, 26-40.

- Solís, M. (2015). The dilemma of combining positive and negative items in scales. *Psicothema*, 27(2), 192-199. <https://doi.org/10.7334/psicothema2014.266>
- Schwartz, S. (2003). *A proposal for measuring value orientations across nations*. En *Questionnaire development report of the european social survey*. Recuperado de <https://bit.ly/2xq7SAg>
- Tardos, A. y Vasseur-Paumelle A. (2018). Reglas y límites en la guardería, adquisición de actitudes. En E. Herrán (Ed.), *Claves de la educación Pikler-Lóczy: Compilación de 20 artículos escritos por sus creadoras* (pp. 377-392). Asociación Pikler-Lóczy de Hungría.
- Tarullo, A., Obradovic, J. y Gunnar, M. (2009). Self-control and the developing brain. *Zero to three*. January, 31-37.
- Thompson, A. G. (1992). Teachers' beliefs and conceptions: A synthesis of the research. En D. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 127-146). Macmillan.
- Tomasello, M. (2007). *Los orígenes culturales de la cognición humana*. Amorrortu.
- Tonyan, H. A., Mamikonian-Zarpas, A. y Chien, D. (2013). Do they practice what they preach? An Ecocultural, multidimensional, group-based examination of the relationship between beliefs and behaviours among child care providers. *Early Child Development and Care*, 183(12), 1853-1877. <https://doi.org/10.1080/03004430.2012.759949>
- Trigwell, K. y Prosser, M. (1996). Changing approaches to teaching: A relational perspective. *Studies in Higher Education*, 21(3), 275-284. <https://doi.org/10.1080/03075079612331381211>
- Vartuli, S. y Rohs, J. (2009). Early childhood prospective teacher pedagogical belief shifts over time. *Journal of Early Childhood Teacher Education*, 30(4), 310-327. <https://doi.org/10.1080/10901020903320262>
- Vincze, M. (2018). La comida del bebé: del biberón a la autonomía En E. Herrán (Ed.), *Claves de la educación Pikler-Lóczy: Compilación de 20 artículos escritos por sus creadoras* (pp. 207-232). Asociación Pikler-Lóczy de Hungría.
- Wallon, H. (1985). *La vida mental*. Crítica.

## Breve Cv de los autores

### Elena Herrán Izagirre

Profesora agregada de la Universidad del País Vasco del departamento de Psicología Evolutiva y de la Educación. Investiga en desarrollo temprano, en intervención educativa en Educación Infantil y en formación del profesorado de esta etapa. En estos ámbitos es autora de varios artículos, ha impartido lecciones y presentado comunicaciones y ponencias en cursos, congresos y jornadas de formación para profesionales de la educación infantil. ORCID ID: <https://orcid.org/0000-0001-8700-6103>. Email: [elena.herran@ehu.eus](mailto:elena.herran@ehu.eus)

### Nuria Galende Pérez

Profesora agregada (contratada doctora) de la Universidad del País Vasco. Trabaja en la Facultad de Educación de Bilbao, en el departamento de Psicología Evolutiva y de la Educación. Imparte docencia e investiga principalmente en áreas relacionadas con el

desarrollo temprano, intervención educativa en Educación Infantil y Primaria, y en formación del profesorado en ambas etapas. En estos ámbitos, es autora de varios artículos y ha participado en diversos congresos nacionales e internacionales. ORCID ID: <https://orcid.org/0000-0002-2195-6912>. Email: [nuria.galende@ehu.eus](mailto:nuria.galende@ehu.eus)

**Gorka Etxebarria Elordui**

Profesor laboral interino de Universidad en la Universidad del País Vasco, en concreto en la Facultad de Educación de Bilbao. Pertenece al departamento de Didáctica y Organización escolar e investiga en desarrollo temprano. Actualmente es doctorando en el programa de Psicodidáctica (Psicología de la Educación y Didácticas Específicas) y ha participado en varios congresos de carácter internacional. ORCID ID: <https://orcid.org/0000-0001-6726-5490>. Email: [gorka.etxebarriae@ehu.eus](mailto:gorka.etxebarriae@ehu.eus)



# Evaluación Auténtica y Evaluación Orientada al Aprendizaje en Educación Superior. Una Revisión en Bases de Datos Internacionales

## Formative Assessment and Learning Oriented Assessment in Higher Education. A Review in International Databases

Emilio José Barrientos-Hernán <sup>1</sup>

Víctor M. López-Pastor \* <sup>2</sup>

Darío Pérez-Brunicardi <sup>2</sup>

<sup>1</sup> Colegio Público Patriarca Eijo Garay, España

<sup>2</sup> Universidad de Valladolid, España

El objetivo de este estudio es realizar un análisis sobre el actual estado de la cuestión en la literatura internacional sobre la temática de la “Evaluación Orientada al Aprendizaje” (*Learning Oriented Assessment*, LOA) y la “Evaluación Auténtica” (*Authentic Assessment*, AA) en Educación Superior (HE). La necesidad surge de los nuevos retos que implica el Espacio Europeo de Educación Superior (EEES) en relación con la Evaluación Formativa y sus distintas formas de entenderla. Para ello se ha llevado a cabo una búsqueda bibliográfica en 3 bases de datos internacionales: WOS, ERIC y *Google Scholar*. Se ha realizado un análisis de contenido a través de un sistema con dos grandes categorías: LOA y AA. Los resultados principales del LOA son: (1) las tareas de evaluación deberían ser diseñadas para estimular el aprendizaje de los alumnos; (2) hay que implicar a los alumnos en la evaluación; (3) el feedback ha de darse en el momento adecuado y debe ser útil para el aprendizaje actual y futuro. La mayoría de experiencias indican aspectos positivos en la utilización de LOA-HE. Los resultados de AA indican que los principales criterios para su definición serían: (1) llevar a cabo tareas auténticas en el aula, con características similares o iguales a las de la vida real; (2) contexto similar o mimetizado al que se encontrarán en el trabajo. La mayoría de experiencias indican que el alumnado valora muy positivamente este tipo de evaluación, por su conexión con la vida real y/o el futuro ambiente de trabajo.

**Palabras clave:** Evaluación auténtica; Evaluación orientada al aprendizaje; Educación superior; Evaluación formativa; Implicación del alumnado en la evaluación.

This paper analyses a review of the international literature on “Learning Oriented Assessment” (LOA) and “Authentic Assessment” (AA) in Higher Education (HE). The European Higher Education Area (EHEA) implies new challenges on Formative Assessment (FA) and the approaches related to it. A literature review has been carried out using three international databases: WOS, ERIC and Google Scholar. A content analysis has been carried out establishing two categories: LOA and AA. The main results on LOA are: (1) Assessment tasks should be designed to stimulate sound learning practices amongst students; (2) Assessment should involve students actively in engaging with criteria, quality, their own and/or peers’ performance; (3) Feedback should be timely and forward-looking so as to support current and future student learning. Most of the experiences show positive aspects using LOA-HE. The results on AA-HE show the main criteria to define it as: (1) do authentic tasks in the classroom with the same or similar characteristics to real life work-tasks; (2) use mimetic or similar contexts to real-work spaces. Most of the experiences show that students assessed very positively this type of assessment due to their connection with real life and/or future world contexts.

**Keywords:** Authentic assessment; Learning oriented assessment; Higher education; Formative assessment; Student involvement in assessment.

---

\*Contacto: vlopez@mpc.uva.es

## 1. Introducción

El Espacio Europeo de Educación Superior (EEES) ha conllevado una serie de adaptaciones en las universidades europeas, cambios legislativos y modificaciones que han repercutido en los roles que juegan profesores, alumnos e instituciones (Bretones, 2008; Pérez y otros, 2008); sus principales principios rectores serían: la renovación de la metodología de enseñanza, la adopción del modelo de enseñanza centrado en el estudiante y la adopción del modelo de enseñanza basado en competencias y la promoción del aprendizaje.

Previamente a la introducción del EEES, un gran número de profesores entendían que la única obligación docente era el ser meros transmisores de conocimientos a sus alumnos; y la manera más habitual de evaluar esos conocimientos era a través de un examen final, que se pudiera traducir en una calificación de la asignatura (Palacios y López-Pastor, 2013). Por ello, algunos autores indicaban que las universidades deberían pasar de una cultura evaluativa basada en el examen a otra centrada en una evaluación para el aprendizaje (Dochy, Segers y Dierick, 2002). En este sentido, el EEES parece implicar que los profesores modifiquen los sistemas de evaluación-calificación tradicionales, dado que la actividad evaluadora ha sido considerada como una de las grandes carencias en las aulas y una de las competencias docentes que más debe cambiar (Zabalza, 2003). Sin embargo, algunas investigaciones recientes muestran cambios en las facultades españolas encargadas de la formación inicial del profesorado (FIP), sobre todo desde el año 2010, donde un tercio del profesorado ya empieza a utilizar sistemas de Evaluación Formativa y Compartida (EFyC) en FIP (Palacios y López-Pastor, 2013). En cambio, otros autores se muestran críticos con este proceso; Bretones (2008) consideraba que las repercusiones de los cambios que se esperaban por la introducción del EEES en España habían sido muy leves y, en la mayoría de los casos, simplemente a nivel burocrático.

Según Palacios y López-Pastor (2013), el sistema de evaluación formativo es de los que mejor se adecua al EEES, ya que su principal finalidad es mejorar el aprendizaje del alumnado y el funcionamiento del proceso de enseñanza y aprendizaje (López-Pastor, 2008 y Pérez y colaboradores, 2008). Por lo tanto, estaría dando respuesta a cómo valorar el proceso del alumno y, a la vez, tener en cuenta el aprendizaje alcanzado por los mismos. No existen excesivos estudios bibliográficos relacionados con la “evaluación formativa en educación superior” (EF en HE) y los realizados son bastante antiguos. Uno de los de mayor impacto fue llevado a cabo por Black y Wiliam (1998), sobre el término: “*Formative Assessment* (FA)”. Para ello recogieron un total de 681 publicaciones relacionadas con la aplicación de un sistema de FA en los colegios y universidades. Entre otros aspectos, resaltaba la reconceptualización del término FA, debido a la confusión que estaba generando, ya que muchas de las publicaciones afirmaban llevar a cabo un sistema de FA pero en la práctica eran similares al sistema de evaluación sumativo. Posteriormente Gaunlett (2007) llevó a cabo una búsqueda bibliográfica del concepto de “FA en *Higher Education* (FA-HE)” entre 1998-2007, seleccionando 46 estudios. Este autor afirma que para que se produzca una evaluación que se pueda definir como formativa hay que considerar el tipo de feedback que se realiza, siendo el *feedback* formativo el elemento clave.

En España, la mayoría de los estudios relacionados con el término de FA-HE comenzaron en el año 2000, pero el mayor incremento ha sido a partir de 2009 (López-Pastor y Sicilia 2017). Barrientos (2013) realiza una búsqueda bibliográfica internacional del concepto de FA-HE entre 2007-2013, así como de otros términos relacionados con el anterior, como:

“*Alternative Assessment*”-ALA (Evaluación Alternativa), “*Authentic Assessment*”-AA (Evaluación Auténtica), “*Assessment For Learning*”-AFL (Evaluación Para el Aprendizaje), “*Learning Oriented Assessment*”-LOA (Evaluación Orientada al Aprendizaje), “*Electronic Assessment*”-EA (Evaluación Electrónica), “*Students Involvement*”-SI: *Peer-Assessment, Self-Assessment* y *Collaborative Assessment* (Implicación del alumnado en su evaluación: evaluación entre iguales, autoevaluación y evaluación colaborativa). Revisando la literatura internacional especializada puede comprobarse que el término FA ha evolucionado hacia otros más centrados en el proceso de aprendizaje del alumnado, como: ALA, AFL, LOA o AA (Carless, 2007; Lorente y Kirk, 2013; López-Pastor y Sicilia, 2017; López-Pastor, Kirk, Lorente-Catalan, Macphail y MacDonald, 2013).

Dos de los conceptos más interesantes de esta red nomológica son los de LOA y AA. De ellos no existen prácticamente estudios y están estrechamente relacionados con el sistema de FA (López-Pastor y Sicilia, 2017). López-Pastor y otros (2013) consideran la AA como un sistema de evaluación estrechamente relacionado con la FA. El concepto de AA hace referencia al hecho de que las técnicas, instrumentos y actividades de evaluación estén claramente aplicados en situaciones, actividades y contenidos reales del aprendizaje. Se opone, por tanto, a las situaciones puntuales y artificiales de evaluación, alejadas de la práctica real o de la aplicación real de dichos conocimientos (Archbald 1991; Archbald y Newman, 1988; Gulikers y otros, 2004; Rule, 2006). El término LOA es utilizado por diferentes autores (Boud y Falchikov, 2006; Carless, 2007; Carmona y Flores, 2008; Ibarra, Rodríguez y Gómez, 2012; Keppell y Carless, 2006) y tiene tres características: (1) el proceso de evaluación debe implicar activamente al estudiante como evaluador; (2) se debe aportar retroalimentación y proalimentación; y (3) las tareas de evaluación deben ser tareas auténticas, ligadas con la realidad profesional (Carless, 2007; Carless, Joughin y Mok, 2006; Ibarra, Rodríguez y Gómez, 2012; Quesada, Rodríguez e Ibarra, 2013).

Según Gessa (2011) el EESS implica un sistema de evaluación en el que se promueva una LOA en el que predominen tareas de carácter auténtico con los alumnos, siendo LOA y AA los dos sistemas que mejor encajarían en esa descripción. Por su parte, Bretones (2008) señala que en España se ha publicado poco sobre LOA y AA en HE.

Debido a la escasez de revisiones bibliográficas realizadas sobre esos conceptos, consideramos interesante y necesario realizar un estudio a fondo del estado de la cuestión de ambos. Por ello, el presente estudio tiene como objetivos:

1. Realizar una búsqueda bibliográfica en tres bases de datos internacionales de los conceptos: “*Learning Oriented Assessment in Higher Education (LOA-HE)*” y “*Authentic Assessment in Higher Education (AA-HE)*”.
2. Analizar el estado de la cuestión de los términos anteriores en la literatura internacional especializada.

## 2. Metodología

Se ha llevado a cabo una búsqueda bibliográfica en tres bases de datos internacionales: “WOS”, “*Google Scholar*” y “ERIC”. Tras la realización de las correspondientes búsquedas y selección de documentos se procedió a realizar un análisis bibliográfico y de contenido de los mismos. Los conceptos de búsqueda han sido: LOA-HE y AA-HE.

Somos conscientes de que dichas búsquedas no agotan todo lo publicado al respecto, porque al limitarnos a estas bases de datos pueden quedarse fuera otros tipos de documentos, como libros y aportaciones a congresos. Se trata de una limitación a asumir en cualquier búsqueda de este tipo.

### 2.1. Técnica e instrumentos de recogida de datos

En cada una de las bases de datos se ha realizado la búsqueda con unos ajustes distintos, ya que cada uno de los buscadores dispone de unos criterios de búsqueda diferentes, siendo imposible realizar la búsqueda con los mismos parámetros en todos los buscadores. El procedimiento fue el siguiente: (a) en *Google Scholar* se ha realizado la búsqueda con los parámetros: “palabras exactas dentro del artículo”; (b) en ERIC se ha analizado: “*topic*” y “*title*”; y (c) en WOS, se ha realizado por: “*topic*” y “*title*”. Por ejemplo, al buscar en *Google Scholar* el concepto de “AA-HE”, hemos recogido información de todos aquellos artículos que muestren dentro del documento ese concepto concreto. En cambio, en ERIC y WOS utilizo como “*title*”: “AA” y como “*topic*”: “HE”.

Del total de publicaciones encontradas, hemos seleccionado aquellas que aportan suficiente información como para poder categorizarlas después. Si la misma publicación aparece en varias bases de datos sólo se selecciona en una de ellas. Se ha realizado la búsqueda con fecha de inicio abierta y hasta mayo de 2018.

Para llevar a cabo la recogida de datos se ha utilizado el cuadro 1, en la que se vuelcan las publicaciones encontradas y seleccionadas en la búsqueda realizada en las tres bases de datos citadas. Para la selección de documentos se han seguido dos criterios: (1) que se cite dentro del documento la frase entera objeto de búsqueda: LOA-HE o AA-HE; y (2) que aporte información relevante de los conceptos de búsqueda: conceptualización, características, clasificaciones, experiencias y/o resultados.

Cuadro 1. Resultados de las búsquedas bibliográficas realizadas

CONCEPTO	BASE DE DATOS	RESULTADOS	SELECCIÓN
Learning Oriented Assessment in Higher Education	WOS	16	3
	ERIC	12	3
	Google Scholar	18	6
Authentic Assessment in Higher Education	WOS	34	7
	ERIC	68	8
	Google Scholar	59	13

Fuente: Elaboración propia.

### 2.2. Técnica de análisis de datos

Para elaborar el informe de resultados y realizar las conclusiones se ha llevado a cabo un proceso de categorización. En el cuadro 2 se han clasificado cada una de las categorías y subcategorías que se han establecido, todas ellas con el objetivo de estructurar y organizar la información recogida.

Cuadro 2. Categorías y subcategorías para analizar los documentos seleccionados

CATEGORÍAS	SUBCATEGORÍAS
1. LOA-HE	1.1- Origen del concepto
	1.2- Criterios para ser LOA
	1.3- Experiencias y resultados de su aplicación
2. AA-HE	2.1- Origen del concepto
	2.2- Criterios para ser AA
	2.3- Experiencias y resultados de su aplicación

Fuente: Elaboración propia.

### 3. Resultados

Los datos están estructurados en base a dos grandes apartados, que obedecen a las categorías y subcategorías establecidas previamente. En las dos categorías se han utilizado cuadros-resumen de los documentos encontrados, ordenando los documentos cronológicamente y aportando una breve explicación del motivo por el que se ha seleccionado ese documento.

#### 3.1. Learning Oriented Assessment in Higher Education

De esta categoría se han seleccionado 11 de los 46 documentos encontrados (cuadro 3).

Cuadro 3. Resultados de la búsqueda del concepto: "LOA-HE"

AUTORES Y AÑO	SELECCIÓN
Boud y Falchikov, 2006	Se explican ejemplos de prácticas de LOA in HE.
Carless, Joughin y Mok, 2006	Se centran en una LOA, dejando de lado y no comparándolo con el de FA, ya que entienden que hay confusión con este término (Yorke, 2003).
Carless, 2007	Introduce el término de LOA y explica sus tres características principales, reflexiona sobre el proyecto llevado a cabo en Hong Kong sobre LOA y explica como la LOA puede ser implementada en la práctica.
Lombard, 2008	Desarrollo del pensamiento crítico a través de una LOA. Además, definen LOA.
Carmona y Flores, 2008	Insisten en la necesidad de que la evaluación se convierta en estrategia para la mejora de los aprendizajes. Para ello, deben cumplirse tres condiciones: 1) las tareas de evaluación deben ser también tareas de aprendizaje; 2) se ha de proporcionar retroalimentación para orientar el trabajo futuro; 3) implicar a los estudiantes en el proceso de evaluar su propio trabajo.
Carless, 2009a	Señala las tres características para ser LOA y analiza las tensiones que genera una evaluación que cumpla con el: "doble deber" que señala Boud.
Carless, 2009b	Uno de los factores que limitan la introducción de prácticas de LOA es la creencia de la poca confianza en las mismas. El autor explica cómo podría ser desarrollada esa confianza.
Carless, 2015a	Selección de 5 profesores premiados por sus buenas prácticas educativas y analizando como sus prácticas evaluativas y su modelo personal de evaluación orientada al aprendizaje.
Carless, 2015b	Propone un modelo de LOA centrada en los tres procesos descritos anteriormente. Además, analizar: 1) las tareas de evaluación que llevan a cabo los alumnos; 2) desarrollo de las capacidades evaluativas de los alumnos; 3) El compromiso de los estudiantes con respecto al feedback.
Rodríguez, Quesada e Ibarra, 2016	Analiza los efectos de una propuesta de LOA-electrónica llevada a cabo por profesores universitarios. Valoran diferentes aspectos de la experiencia implementada.
Canabal y Margalef, 2017	Analiza los procesos de Feedback y algunos instrumentos de evaluación desde una perspectiva centrada en LOA.

Fuente: Elaboración propia.

#### 3.2. Origen del concepto LOA

Según Boud y Falchikov (2006), la LOA es un sistema de evaluación que no se ciñe únicamente al contexto del aula y debe ser útil para los alumnos a la hora de contextualizar lo aprendido en la esfera de la vida real y del trabajo. Por su parte, Carless, Joughin y Mok (2006) señalan la LOA como un concepto no comparable con el de FA, ya que entienden que hay confusión con este término. Por lo que, el término LOA surge para dar respuesta

al doble deber que ha de cumplir la evaluación en HE (Boud y Falchikov, 2006), teniendo que aunar tanto las características de la evaluación sumativa o certificación del alumno como las de la FA o progreso de aprendizaje del alumnado (Carless, 2007, 2009, 2015a).

### ***3.3. Criterios para ser LOA***

La LOA se caracteriza por tres principios (Carless 2007, 2009a, 2015a; Carless, Joughin y Mok, 2006; Carmona y Flores, 2008): (1) las tareas de evaluación deberían ser diseñadas para estimular el aprendizaje de los alumnos; (2) la evaluación tiene que implicar a los alumnos en su propia evaluación y en la evaluación de otros compañeros; y (3) el feedback o retroalimentación ha de darse en el momento adecuado a los alumnos y que sea útil para el aprendizaje actual y futuro de los mismos. Rodríguez, Quesada y Ibarra (2016) denominan “LOA Electrónica” a una reformulación de los tres principios de LOA, pero concretados en un contexto de evaluación electrónica: (1) llevar a cabo tareas de evaluación electrónica como tareas de aprendizaje; (2) utilizar el e-feedback (retroalimentación electrónica) como feedforward (retroalimentación) que permita avanzar hacia delante, en este caso permitiendo a los alumnos orientar su aprendizaje actual y futuro; y (3) participación de los alumnos en la evaluación electrónica a través de estrategias como autoevaluación, evaluación entre iguales y evaluación colaborativa.

Lombard (2008) le da un cariz diferente a la LOA, focalizada en desarrollar el pensamiento crítico de los alumnos. Además, señalan que LOA puede ser descrita como una sinergia entre instrucción, aprendizaje, evaluación y feedback aportado durante la clase, la cual de manera consciente intenta ayudar a los aprendices para apoyar y estimular su competencia en pensamiento crítico.

### ***3.4. Experiencias y resultados de la aplicación LOA***

Carless (2007) explica la primera experiencia en la que se pone en práctica la LOA, en el Instituto de Educación de Hong Kong, a lo largo de cuatro años (2002-2006), en el que se involucró a otras universidades tanto nacionales como extranjeras. El principal objetivo era identificar, promover y divulgar buenas prácticas de LOA-HE. Desde el proyecto se invitó a casi 400 profesores a implementar LOA en sus aulas, de los que 40 realizaron un informe final en el que exponían sus resultados. También realizó un programa de formación inicial con 35 profesores, con una duración de 12 semanas (30 horas). Este programa estaba dirigido a que los profesores conocieran y desarrollaran algunas de las mejores y más innovadoras prácticas docentes realizadas utilizando LOA; que fue valorado de manera muy positiva.

Carless (2015b) realiza una revisión de la bibliografía sobre evaluación en HE, concluyendo en los tres principios de la LOA explicados anteriormente. A partir de estos principios, proponen dos preguntas que han de ser contestadas a lo largo de su estudio: (1) ¿Cómo es la LOA que llevan a cabo una muestra de profesores premiados en sus prácticas docentes?; y (2) ¿Cuáles son las principales percepciones de profesores y alumnos sobre el uso de la LOA? En los resultados hablan de casos concretos: por ejemplo, en el caso de un profesor de arquitectura, son varios los factores que mejoran y facilitan los procesos evaluativos: el uso del portafolio como actividad de evaluación, los procesos de autoevaluación generados en la revisión de los trabajos de los estudiantes y el feedback dialógico creado durante todo el proceso en, las interacciones entre profesores y estudiantes. En otros casos, las clases de historia y derecho eran muy numerosas, pero ello no representó una barrera a la hora de llevar a cabo prácticas de LOA; por lo que los

autores consideran que estas experiencias muestran que la determinación y compromiso de los profesores puede derribar algunos de los problemas que puedan surgir para desarrollar procesos de evaluación efectivos.

Rodríguez y otros (2016) analizan los efectos de la puesta en práctica de un programa de formación dirigido a profesores universitarios de "LOA electrónica". Los resultados muestran que el profesorado que usa herramientas electrónicas de LOA mejora la competencia y habilidad en evaluación, mencionando que se sienten más preparados para llevar a cabo procesos de LOA electrónica en sus prácticas docentes, tienen más recursos bibliográficos, conocen fuentes de información relacionadas con LOA y son capaces de utilizar y poner en práctica procesos de evaluación colaborativa y participativa en sus prácticas docentes.

Canabal y Margalef (2017) llevan a cabo un análisis de los procesos de feedback que se generan entre estudiantes y profesores desde una perspectiva de LOA. El contexto de la investigación es el Máster en Docencia Universitaria, en el que participaron profesores de distintas disciplinas de la universidad. Los participantes señalaron que el proceso de elaboración de cartas personales como instrumentos de feedback aumentó la motivación y aprendizaje de los participantes. Además, en este estudio se clasificaron y analizaron las condiciones en la que los procesos de feedback tienen mayor impacto en una LOA.

Sin embargo, hay varios documentos que citan algunas dificultades para la introducción de sistemas LOA. Carless (2009a) señala que el doble deber que tiene que cumplir la evaluación ha generado una serie de tensiones entre profesores y alumnos, proponiendo una serie de estrategias de evaluación que los estudios han indicado como efectivas, entre las que se encuentran el uso del portafolio (*portfolio*) y de trozos de texto con las claves de los aspectos realizados (*patchwork texts*). Por otro lado, Carless (2009b) analiza como la desconfianza es uno de los factores que incide para en la no utilización de sistemas de evaluación alternativos y LOA. Explica cómo ha de ser desarrollada la confianza, algunas de las barreras que hay que eliminar y la relación entre confianza y buenas prácticas evaluativas. Algunas de las estrategias que propone para crear confianza son: (1) sistemas de evaluación que puedan ser justificados de manera teórica y práctica, que además aporten la seguridad para defender nuestras prácticas evaluativas frente a evaluadores internos y externos; y (2) gran transparencia en los procesos de evaluación.

Los principales resultados de esta primera categoría podrían ser los siguientes: (1) Tanto el origen, los criterios y la mayoría de los trabajos sobre la LOA parecen estar ligados al autor Carless (2006, 2007, 2009b, 2015), aunque también aparecen otros autores que aportan trabajos e ideas novedosas en algunas temáticas; y (2) los resultados parecen ser positivos en la mayoría de los estudios realizados aunque también surgen problemas y resistencias, como en cualquier tipo de innovación educativa. Alguno de los estudios aporta algunas posibles estrategias para minimizar o solucionar dichas dificultades.

### 3.5. Authentic Assessment in Higher Education

El segundo concepto de análisis es el de AA in HE; se han seleccionado 28 de los 161 documentos encontrados (cuadro 4).

Cuadro 4. Resultados de la búsqueda del concepto: "AA-HE"

AUTORES Y AÑO	SELECCIÓN
Archbald y Newman, 1988	Origen del término AA del que parten la mayoría de estudios posteriores.

Archbald, 1991	Explica los principios, experiencias y temas relacionados con AA-HE, todo ello relacionado con el mundo del trabajo
Cumming y Maxwell, 1999	Concepto relacionado con el mundo del trabajo y clasifican las propiedades de la AA en: desempeño, contexto, complejidad y competencia.
Gulikers, Bastiaens y Kirshner, 2004	Describen un marco de referencia de la AA con cinco dimensiones para que pueda tener lugar: (1) la tarea de evaluación; (2) el contexto físico de la tarea; (3) el contexto social de la tarea; (4) la evaluación de resultados; (5) los criterios de evaluación.
Rule, 2006	Revisión de artículos de AA-HE, aunando en cuatro las características de las que se deben componer las actividades de AA: (1) involucrar problemas del mundo real que mimeticen situaciones de trabajo de los profesionales de esa materia; (2) incluir situaciones con respuestas abiertas y múltiples opciones, habilidades de reflexión y metacognición; (3) involucrar a los alumnos en debates y en el aprendizaje de habilidades sociales; (4) permitir a los estudiantes dirigir su propio aprendizaje.
Gulikers, Kester, Kirschner y Bastiaens, 2008	Estudio en el que analizan cómo los estudiantes perciben la AA y cómo influye en sus aprendizajes.
Keyser y Howell, 2008	Realizan una revisión de artículos relacionados con AA-HE.
Hassanpour, Utaberta, Abdullah y Tahir, 2011	AA como una evaluación diferente a la tradicional o sumativa, dándole importancia no sólo a adquirir conocimientos.
Vu, 2011	Relaciona la AA con el mundo de la educación formal.
Bohemia y Davison, 2012	Relación de AA con el mundo del trabajo
Eddy y Lawrence, 2013	Aportan cuatro características de las que se tiene que componer la AA: (1) la evaluación es un proceso y no es algo estático y puntual; (2) la AA supone evaluar aprendizajes experimentales; (3) que sean varias las personas que evalúen el trabajo del estudiante, incluyendo la auto-evaluación o la revisión por una audiencia pública; y (4) la AA tiene que ofrecer más oportunidades al aprendiz para decidir en su evaluación.
Bosco y Ferns, 2014	Clasifican las actividades de AA en base a dos criterios: (1) autenticidad; y (2) proximidad.
Ashford-Rowe, Herrington y Brown, 2014	Realizan una revisión bibliográfica y reformulan en ocho preguntas las características que deben tener las prácticas de AA.
Biddle, 2014	Desde el programa de doctorado de la Universidad Central de Florida llevan a cabo en colegios y distritos varios proyectos relacionados con AA.
Gonzalez-DeHass y Willems, 2015	Cursos de formación en la facultad de psicología con futuros profesores, en los que utilizan ambientes de aprendizaje auténtico y como llevar a cabo una AA.
Heinzen, Landrum, Gurung y Dunn, 2015	Señalan las ocho grandes dificultades encontradas para implementar AA-HE.
Latorre y Varela, 2015	Uso de las rúbricas para que aporten feedback y feedforward.
Kearney, Perkins y Kennedy-Clark, 2016	Ponen en práctica una AA en la que los alumnos realizan autoevaluaciones y evaluaciones a otros compañeros bastante precisas.
Sullivan y McConnell, 2017	Las universidades deben actualizarse al siglo XXI, entre otras cosas introduciendo una AA.
Prieto, Llácer y Escobar, 2017	Utilizan el portafolio como uno de los instrumentos más populares para pasar de una evaluación tradicional a otra basada en AA
Ghosh, Bowles, Ranmuthugala y Brooks, 2017	Realizan una revisión de estudios de AA, concluyendo que es necesario que estas prácticas cumplan aspectos como fiabilidad y validez.
Febriana y Arlianty, 2017	Desde AA, desarrollan la incidencia de una herramienta que sea capaz de evaluar: conocimientos, actitudes, auto-eficacia y evaluación entre iguales. Facultad de Química en Indonesia.

McDermott y otros, 2017	Utilización de una AA que este en consonancia con la metodología que se lleva a cabo.
Santos, 2017	Pone en práctica una AA en la Facultad de Farmacia en la que los estudiantes encuentran difícil autoevaluarse o evaluar a otros compañeros,
Murphy, Fox, Freeman y Hughes, 2017	Aportan una definición de AA y una guía para su implementación en el aula
Kaider, Hains-Wesson y Young, 2017	Clasifican las actividades de AA en base a dos criterios: autenticidad y proximidad.
Nguyen, 2017	Estudio en Vietnam en el que los futuros profesores introducen tres tareas de AA en sus prácticas.
James y Casidy, 2018	Aplican AA en una muestra de 120 alumnos en la Facultad de Negocios con resultados positivos.

Fuente: Elaboración propia.

### 3.6. Origen del concepto AA

Archbald (1991) fue uno de los primeros autores que definió el término de AA-HE como una evaluación utilizada para evaluar experiencias y logros del mundo real. Además, señalan que la mejor manera de evaluar el aprendizaje que ocurre en el aula clase es aquella realizada de manera auténtica (Archbald y Newman, 1988; Keyser y Howell, 2008). Cumming y Maxwell (1999) añaden que la aplicación de AA variará dependiendo de las creencias educativas de quién lo lleva a la práctica, indicando que la AA es una forma coherente de llevar a cabo la evaluación, siempre y cuando esté en consonancia los procesos de enseñanza y aprendizaje.

Muchos estudios relacionan el concepto AA con el mundo del trabajo (Archbald, 1991; Archbald y Newman, 1988; Bohemia y Davison, 2012; Cumming y Maxwell, 1999). Vu (2011) añade que la AA implica unos fuertes lazos con el mundo de la educación formal, centrando su propuesta con el mundo educativo, dirigida a promover los aprendizajes de los estudiantes y con un carácter más humanizador y que sirva para preparar a los estudiantes para el futuro; además de evaluar los aprendizajes, el alumnado también interioriza sus conocimientos y aprende como actuar con las personas que tiene alrededor. Murphy y otros (2017) añaden que la AA es un tipo de evaluación en la cual los estudiantes tienen que desarrollar tareas del mundo real en las que demuestren la aplicación de conocimientos esenciales y habilidades clave.

En la misma línea educativa, Hassanpour y otros (2011) definen AA como una evaluación que no está solo centrada en la adquisición de conocimientos como fin último, sino también en evaluar durante el proceso el uso de conocimientos, habilidades y estrategias para la resolución de problemas con múltiples opciones de solución. Además, utilizan la AA como sinónimo de una evaluación alternativa y diferente a la que ellos denominan “estándar o tradicional”. En línea con lo anterior, Sullivan y Mcconnell (2017) indican que las facultades universitarias deben proponerse una serie de retos que estén en concordancia con las necesidades y expectativas del estudiante universitario del siglo XXI. Uno de ellos es llevar a cabo una AA entendida como una evaluación alternativa y/o FA que este en coherencia con la metodología utilizada. McDermott y colaboradores (2017) señalan la necesidad de llevar a cabo un tipo de AA que este en concordancia con metodologías como el aprendizaje situado y aprendizaje social, señalando la importancia de una evaluación que guie el aprendizaje.

### **3.7. Criterios para ser AA**

Cumming y Maxwell (1999) clasificaron la AA y sus propiedades en cuatro subgrupos: desempeño, contexto, complejidad y competencia. Además, asociaron las tres primeras con teorías del aprendizaje: (1) “desempeño”; similitud entre las tareas de evaluación y las de la vida real; (2) “contexto”; los estudiantes suelen transferir muy pocas de las destrezas aprendidas en el aula a situaciones del mundo real; se relaciona con la teoría del aprendizaje situado y la necesidad de utilizar contextos de aprendizaje lo más reales posibles; (3) “complejidad”; los estudiantes están mejor preparados para desarrollar la capacidad de resolución de problemas si el aprendizaje y la evaluación están inherentes en escenarios complejos que simulen situaciones originales y genuinas; (4) “competencia”; actividades que requieren del uso de diferentes habilidades para su desempeño.

Para Gulikers y colaboradores (2004), la AA se distingue por tener cinco dimensiones: (1) la tarea de evaluación; (2) el contexto físico de la tarea; (3) el contexto social de la tarea; (4) la evaluación de resultados; y (5) los criterios de evaluación. Además, sostienen que las tareas llevadas a cabo deben reflejar la competencia que necesita ser evaluada, su contenido tiene que representar situaciones o problemas de la vida real y los estudiantes deberían llevar a cabo procesos de reflexión similar a los utilizados para solucionar ese mismo de situaciones cuando ocurren en la vida real.

Rule (2006) realiza una revisión de ejemplos de AA en HE y señala cuatro características de las que se deben componer las actividades de AA: (1) involucrar problemas del mundo real que mimeticen situaciones de trabajo de los profesionales de esa materia; (2) incluir situaciones con respuestas abiertas y múltiples opciones, habilidades de reflexión y metacognición; (3) involucrar a los alumnos en debates y en el aprendizaje de habilidades sociales; y (4) permitir a los estudiantes dirigir su propio aprendizaje.

Por otra parte, algunos estudios analizan las actividades de AA en base a dos aspectos (Bosco y Ferns, 2014; Kaider, Hains-Wesson y Young, 2017): (1) “autenticidad”; siendo más genuina cuanto más parecida son las habilidades necesarias para desarrollar la tarea simulada a la que se encontraría en la realidad; y (2) “proximidad”; siendo mayor cuando el contexto reúne las mismas características o similares a las encontradas en la realidad.

Eddy y Lawrence (2013) presentan una estructura conceptual de cuatro fases en AA: (1) la evaluación es un proceso y no es algo estático y puntual; (2) la AA supone evaluar aprendizajes experimentales; (3) que sean varias las personas que evalúen el trabajo del estudiante, incluyendo la auto-evaluación o la revisión por una audiencia pública; y (4) la AA tiene que ofrecer más oportunidades al aprendiz para decidir en su evaluación.

Ashford-Rowe, Herrington y Brown (2014) realizan una revisión bibliográfica y establecen ocho preguntas que deben caracterizar las prácticas de AA: ¿Hasta qué punto la actividad evaluadora le supone un reto al alumno? ¿Es el proceso o el resultado el objetivo final de la evaluación? ¿Requiere la evaluación de una transferencia del aprendizaje haciendo uso de las habilidades aprendidas? ¿Requiere la evaluación de procesos metacognitivos? ¿El resultado o proceso de la evaluación puede ser reconocida como auténtica por los partícipes de la misma?

¿Son los instrumentos de evaluación reales o simulados? ¿Requiere la evaluación procesos de discusión y retroalimentación? ¿Requiere la evaluación que los alumnos colaboren entre ellos?

Murphy y otros (2017) realizan una guía con una serie de fases y pasos para que el profesorado universitario puede implementar una AA en su proceso de enseñanza, estableciendo cinco fases: (1) identificación de los objetivos de aprendizaje que se desean y alinearlos con las tareas que se propongan; (2) generar procesos de comunicación y consulta con los estudiantes; (3) desarrollo de rúbricas y criterios de evaluación; (4) implementación tareas de AA y mejorar el proceso aportando feedback formativo; y (5) evaluar y reflexionar sobre lo ocurrido en el proceso de evaluación.

### ***3.8. Experiencias y resultados de la aplicación AA***

Gulikers y otros (2008) realizan un estudio en el que preguntan cómo perciben los alumnos la AA y cómo condiciona esta sus aprendizajes. Sus resultados muestran que los estudiantes muy experimentados perciben mejor la influencia de la AA en el aprendizaje que los poco experimentados; sugiriendo posibles líneas de acción para desarrollar y usar la AA.

Son muchos los estudios que utilizan las características de la AA en experiencias reales. Desde el programa de doctorado de la Universidad Central de Florida llevan a cabo varios proyectos en los que utilizan una AA en los colegios y distritos de la zona; en los que los estudiantes del doctorado trabajan directamente con el colegio o distrito (Biddle, 2014). Los resultados señalan que los alumnos necesitan conocer la situación real de su futuro trabajo, conocer e interiorizar las necesidades de la organización y aprender las habilidades necesarias para desarrollar ese trabajo. Gonzalez-DeHass y Willems (2015), en unos cursos de psicología con alumnos que serán futuros profesores, desarrollan ambientes de aprendizaje auténtico y prácticas reales de cómo llevar a cabo una AA. Los resultados indican que es positivo que los maestros puedan aplicar sus aprendizajes en contextos reales y enfrentarse con experiencias reales en las que tengan que tomar sus propias decisiones.

Son varios los estudios que utilizan las nuevas tecnologías para llevar a cabo una AA. Eddy y Lawrence (2013) utilizan las Wikis como plataformas web en las que se produce a la vez aprendizaje y AA. Como resultados señalan que se producen aprendizajes auténticos y no es necesaria ninguna formación previa relacionada con la informática o la programación para realizar las actividades. Heinzen, Landrum, Gurung y Dunn (2015) consideran que llevada a cabo AA en HE es un reto difícil y señalan ocho grandes dificultades, utilizando la Gamificación para darles una posible solución, encontrando una motivación y actitud positiva de los estudiantes hacia este formato.

Hay varios estudios que utilizan la rúbrica y portafolio como instrumentos de evaluación imprescindibles para llevar a cabo una AA. Keyser y Howell (2008) llevan a cabo una revisión de artículos relacionados con la AA-HE, concluyendo que si los modelos teóricos que definen y describen la AA son coherentemente evaluados por rúbricas diseñadas en consonancia con los objetivos de aprendizaje, entonces se producirán no solo una evaluación eficiente, sino el desarrollo de mejores aprendizajes. Latorre y Varela (2015) señalan que las rúbricas proporcionan al estudiante una doble vía de información: le aporta una mayor información que una simple valoración de su trabajo (entendido como feedback) y le guía sobre cómo mejorar los logros alcanzados, generando un diálogo con el profesor y/o sus compañeros sobre los aspectos que debe acometer en el futuro (entendido como feedforward). Sullivan y McConnell (2017) piensan que crear rúbricas que evalúen aspectos claves del aprendizaje puede ser una herramienta útil para estructurar los tiempos en el proceso de enseñanza y cuando deben ser introducidas nuevas tareas de

aprendizaje a los alumnos. Respecto al uso del portafolio, Prieto, Llacer y Escobar (2017) indican que es uno de los instrumentos más populares para pasar de una evaluación tradicional a otra basada en AA, ya que evalúa el desempeño de los alumnos y también mejora su aprendizaje y ayuda a que reflexionen sobre el mismo.

Otro de los aspectos tratados habitualmente es que la evaluación no sea realizada únicamente por el profesor y que puedan participar los alumnos. Kearney, Perkins y Kennedy-Clark (2016) muestran que los estudiantes son capaces de juzgar su propio trabajo y realizar evaluaciones bastante precisas del trabajo de sus compañeros, incluso aquellos que no tenían experiencia previa en autoevaluarse o en llevar a cabo una evaluación entre iguales. Sin embargo, Santos (2017) lleva a cabo una AA en la Facultad de Farmacia en donde los estudiantes encuentran difícil evaluar sus propios trabajos o los de otros compañeros al mismo nivel que lo realizan los profesores.

Febriana y Arlianty (2017) analizan en la Facultad de Química en Indonesia el posible efecto de la aplicación de AA en los futuros profesores. La herramienta de AA utilizada consistía en revisar aspectos de conocimientos, actitudes, auto-eficacia y evaluación entre iguales. Los resultados muestran mejoría en todos los aspectos señalados anteriormente. Por su parte, James y Casidy (2018) señalan que la aplicación de una AA en la Facultad de Negocios genera actitud y comportamientos positivos hacia las tareas propuestas. Por último, Nguyen (2017) lleva a cabo un estudio en un módulo sobre pedagogía en Vietnam en FIP. Para ello se implementan tres tareas de AA en la que los futuros maestros perciben que la AA influyó y cambió sus estrategias de aprendizaje y motivación, sus competencias profesionales y su identidad como profesores.

Ghosh y otros (2017) realizan una revisión de estudios AA, señalando que la tónica general es la falta de validez y fiabilidad de los mismos. Por ello, piensan que, si los aspectos de validez y fiabilidad de las prácticas AA son mejoradas desde una visión holística de la realidad, la evaluación del programa y el desarrollo de tareas por parte de los estudiantes en el lugar de trabajo pueden mejorar significativamente.

Por tanto, parecen existir orígenes diferentes del concepto “AA”, desde diferentes campos profesionales y educativos. Esto hace que también podamos encontrar diferentes teorías y planteamientos sobre que es el AA y que criterios debe cumplir un sistema de evaluación para ser denominado así. A pesar de ello, si pueden encontrarse algunos criterios comunes para definir que es AA, que serán presentados en el apartado de conclusiones. La mayoría de los estudios realizados muestra resultados positivos cuando se utilizan sistemas o situaciones de AA en HE, aunque también existen trabajos que cuestionan la validez y fiabilidad de la mayoría de los estudios realizados sobre esta temática. Por otra parte, parece que los instrumentos de evaluación más adecuados y utilizados para desarrollar procesos de AA en HE son el portafolios y las rúbricas.

## **4. Conclusiones**

A lo largo de este estudio hemos realizado una revisión bibliográfica en tres bases de datos internacionales de los conceptos: LOA-HE y AA-HE. Se han encontrado un total de 208 documentos de los que se han seleccionado 40. Del término LOA-HE se han encontrado 46 documentos y se han seleccionado 12, todos ellos están centrados en el ámbito de la enseñanza. Además, hay un gran número de referencias en las que Carless aparece como autor principal y/o es citado por otros autores. En cuanto al concepto de AA-HE se han

encontrado 161 documentos y se han seleccionado 28, un número considerablemente más alto, que muestra una mayor presencia en la literatura internacional, en parte por la mayor relación que tiene con el mundo del trabajo.

Los resultados muestran que el término LOA-HE surge inicialmente de la necesidad de responder al doble deber de la evaluación en HE: potenciar el aprendizaje del alumnado y asegurar la certificación al final de una asignatura y/o curso. Además, la LOA se intenta distanciar del término FA, ya que señalan que este último está rodeado de cierta confusión. La mayoría de los documentos señalan las características que cita Carless en sus estudios como propias de una LOA-HE: (1) las tareas de evaluación deberían ser diseñadas para estimular el aprendizaje de los alumnos; (2) la evaluación tiene que implicar a los alumnos en su propia evaluación y en la evaluación de otros compañeros; y (3) el feedback o retroalimentación ha de darse en el momento adecuado a los alumnos y que sea útil para el aprendizaje actual y futuro de los mismos. La mayoría de experiencias indican aspectos positivos en la introducción de LOA-HE en sus programas formativos en educación superior.

Por otra parte, los resultados muestran que el origen del término AA-HE se remonta al mundo del trabajo y a la necesidad de llevar a cabo en el aula tareas lo más similares posibles a las condiciones que se encontrarán los alumnos en sus futuros ambientes de trabajo. Sin embargo, hay otros documentos que señalan las virtudes de poder llevar a cabo una AA con una finalidad educativa, en la que los dos grandes instrumentos de evaluación son el diseño de rúbricas y el portafolio. Además, un gran número de documentos señalan unos criterios comunes para llevar a cabo una AA: (1) llevar a cabo tareas auténticas en el aula, con características similares o iguales a las de la vida real; y (2) contexto similar o mimetizado al que se encontrarán en el trabajo. La mayoría de experiencias indican que los alumnos que reciben este tipo de evaluación la valoran muy positivamente, especialmente la conexión con su futuro laboral, las habilidades requeridas y la puesta en práctica en situaciones reales.

Este estudio es clave en el estado actual de la evaluación en el EEES. La mayoría de revisiones bibliográficas relacionadas con FA-HE están hechas en inglés, poco actualizadas y sin considerar los conceptos de LOA-HE y AA-HE, que son muy utilizados actualmente en la literatura internacional. Por ello, este trabajo presenta una clarificación terminológica en castellano y las características básicas de cada concepto, que son congruentes con las exigencias que implica el EEES. Por todo ello puede resultar un trabajo muy útil para todo el profesorado universitario interesado en estas temáticas, así como los gabinetes de evaluación y mejora de las universidades iberoamericanas.

Como prospectiva, sería interesante investigar si se están publicando en español experiencias y estudios relacionadas con LOA-HE y AA-HE y que resultados están obteniendo; así como la posible comparativa de estos resultados con los encontrados a nivel internacional.

## Referencias

- Archbald, D. A. (1991). Authentic assessment: principles, practices, and issues. *School Psychology Quarterly*, 6(4), 279-293. <https://doi.org/10.1037/h0088821>
- Archbald, D. A. y Newmann, F. M. (1988). *Assessing authentic academic achievement in the secondary school*. NASSP.

- Ashford-Rowe, K., Herrington, J. y Brown, C. (2014). Establishing the critical elements that determine authentic assessment. *Assessment & Evaluation in Higher Education*, 39(2), 205-222. <https://doi.org/10.1080/02602938.2013.819566>
- Barrientos, E. (2013). *La evaluación formativa y evaluación orientada al aprendizaje en educación superior: una revisión internacional*. Universidad de Valladolid.
- Biddle, J. (2014). *A needs analysis for K-12 school improvement projects and their use as the dissertation in practice for the professional practice education doctorate program at the University of Central Florida*. University of Central Florida.
- Black, P. y Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, 5(1), 7-74. <https://doi.org/10.1080/0969595980050102>
- Bohemia, E. y Davison, G. (2012). Authentic learning: the gift project. *Design and Technology Education: an International Journal*, 17(2), 49-61.
- Bosco, A. M. y Ferns, S. (2014). Embedding of authentic assessment in work-integrated learning curriculum. *Asia-Pacific Journal of Cooperative Education*, 15(4), 281-290.
- Boud, D. y Falchikov, N. (2006). Aligning assessment with long-term learning. *Assessment y Evaluation in Higher Education*, 25(3), 279-291. <https://doi.org/10.1080/02602930600679050>
- Bretones, A. (2008). Participación del alumnado de Educación Superior en su evaluación. *Revista de Educación*, 347, 181-202.
- Canabal, C. y Margalef, L. (2017). The feedback: a key to learning-oriented assessment. *Profesorado. Revista de Curriculum y Formación de Profesorado*, 21(2), 149-170.
- Carless, D. (2007). Learning-oriented assessment: conceptual bases and practical implications. *Innovations in Education and Teaching International*, 44(1), 57-66. <https://doi.org/10.1080/14703290601081332>
- Carless, D. (2009a). Learning-oriented assessment: principles, practice and a project. En L. H. Meyer, S. Davidson, H. Anderson, R. Fletcher, P. M. Johnston y M. Rees (Eds.), *Tertiary assessment y higher education student outcomes: Police, practice & research learning-oriented assessment* (pp. 79-90). Ako Aotearoa.
- Carless, D. (2009b). Trust, distrust and their impact on assessment reform. *Assessment & Evaluation in Higher Education*, 34(1), 79-89. <https://doi.org/10.1080/02602930801895786>
- Carless, D. (2015a). *Excellence in university assessment: learning from award-winning practice*. Routledge.
- Carless, D. (2015b). Exploring learning-oriented assessment processes. *Higher Education*, 69(6), 963-976. <https://doi.org/10.1007/s10734-014-9816-z>
- Carless, D., Joughin, G. y Mok, M. (2006). Learning-oriented assessment: principles and practice. *Assessment & Evaluation in Higher Education*, 31(4), 395-398. <https://doi.org/10.1080/02602930600679043>
- Carmona, M. y Flores, J. (2008). Learning-oriented assessment in higher education: conditions and strategies for its application to university teaching. *Revista Española de Pedagogía*, 241, 467-486.
- Cumming, J. y Maxwell, G. S. (1999). Contextualising authentic assessment. *Assessment in Education: Principles, Policy & Practice*, 6(2), 177-194. <https://doi.org/10.1080/09695949992865>

- Dochy, F., Segers, M. y Dierick, S. (2002). Nuevas vías de aprendizaje y enseñanza y sus consecuencias: una nueva era de evaluación. *Revista de Docencia Universitaria*, 2(2), 13-29.
- Eddy, P. L. y Lawrence, A. (2013). Wikis as platforms for authentic assessment. *Innovative Higher Education*, 38(4), 253-265. <https://doi.org/10.1007/s10755-012-9239-7>.
- Febriana, B. y Arlianty, W. (2017) The application of authentic assessment in chemistry curriculum. En B. Budi, L. Kim, Y. Isao e I. Khan (Eds.), *Proceedings of the 3rd international conference on education and training* (pp. 99-105). International Institute of Knowledge Management. <https://doi.org/10.17501/icedu.2017.3111>
- Gauntlett, N. (2007). *Literature review on formative assessment in higher education*. Middlesex University.
- Gessa Perera, A. (2011). La coevaluación como metodología complementaria de la evaluación del aprendizaje. Análisis y reflexión en las aulas universitarias. *Revista de Educación*, 354, 749-764.
- Ghosh, S., Bowles, M., Ranmuthugala, D. y Brooks, B. (2017). Improving the validity and reliability of authentic assessment in seafarer education and training: a conceptual and practical framework to enhance resulting assessment outcomes. *WMU Journal of Maritime Affairs*, 16(3), 455-472. <https://doi.org/10.1007/s13437-017-0129-9>
- Gonzalez-DeHass, A. R. y Willems, P. P. (2015). Case-study instruction in educational psychology: implications for Teacher preparation. En M. Li y Y. Zhao (Ed.), *Exploring learning & teaching in higher education* (pp. 99-122). Springer.
- Gulikers, J. M., Bastiaens, T. J. y Kirshner, P. A. (2004). A five-dimensional framework for authentic assessment. *Educational Technology Research and Development*, 52(3), 67-87. <https://doi.org/10.1007/BF02504676>
- Gulikers, J. T., Kester, L., Kirschner, P. A. y Bastiaens, T. J. (2008). The effect of practical experience on perceptions of assessment authenticity, study approach, and learning outcomes. *Learning and Instruction*, 18(2), 172-186. <https://doi.org/10.1016/j.learninstruc.2007.02.012>
- Hassanpour, B., Utaberta, N., Abdullah, N. y Tahir, M. (2011). Authentic assessments or standardized assessment new attitude to architecture assessment. *Procedia-Social and Behavioral Sciences*, 15, 3590-3595. <https://doi.org/10.1016/j.sbspro.2011.04.340>
- Heinzen, T. E., Landrum, R. E., Gurung, R. A. y Dunn, D. S. (2015). Game-based assessment: the mash-up we've been waiting for. En T. Reiners y L. Wood (Eds.), *Gamification in education and business* (pp. 201-217). Springer.
- Ibarra, M. S., Rodríguez, G. y Gómez, M. A. (2012). La evaluación entre iguales: beneficios y estrategias para su práctica en la universidad. *Revista de Educación*, 359, 206-231. <https://doi.org/10.4438/1988-592X-RE-2010-359-092>
- James, L. T. y Casidy, R. (2018). Authentic assessment in business education: its effects on student satisfaction and promoting behaviour. *Studies in Higher Education*, 43(3), 401-415. <https://doi.org/10.1080/03075079.2016.1165659>
- Kaider, F., Hains-Wesson, R. y Young, K. (2017). Practical typology of authentic work-integrated learning activities and assessments. *Asia-Pacific Journal of Cooperative Education*, 18(2), 153-165.
- Kearney, S., Perkins, T. y Kennedy-Clark, S. (2016). Using self-and peer-assessments for summative purposes: analysing the relative validity of the AASL (Authentic Assessment for Sustainable Learning) model. *Assessment & Evaluation in Higher Education*, 41(6), 840-853. <https://doi.org/10.1080/02602938.2015.1039484>

- Keppell, M. y Carless, D. (2006). Learning-oriented assessment: a technology-based case study. *Assessment in Education: Principles, Policy & Practice*, 13(2), 179-191. <https://doi.org/10.1080/09695940600703944>
- Keyser, S. y Howell, S. L. (2008). *The state of authentic assessment*. Columbus.
- Latorre, M. A. y Varela, J. L. M. (2015). La contribución de las rúbricas a la práctica de la evaluación auténtica. *Observar*, 9, 5-17.
- Lombard, B. (2008). Modelling critical thinking through learning-oriented assessment. *South African Journal of Higher Education*, 22(5), 1029-1043. <https://doi.org/10.4314/sajhe.v22i5.42923>
- López-Pastor, V. M. (2008). Implementing a formative and shared assessment system in higher education teaching. *European Journal of Teacher Education*, 31(3), 293-311. <https://doi.org/10.1080/02619760802208452>
- López-Pastor, V. M., D. Kirk, E. Lorente-Catalán, MacPhail, A. y Macdonald, D. (2013). Alternative assessment in physical education: a review of international literature. *Sport, Education & Society*, 18(1), 57-76. <https://doi.org/10.1080/13573322.2012.713860>
- López-Pastor, V. y Sicilia-Camacho, A. (2017). Formative and shared assessment in higher education. Lessons learned and challenges for the future. *Assessment & Evaluation in Higher Education*, 42(1), 77-97. <https://doi.org/10.1080/02602938.2015.1083535>
- Lorente, E. y Kirk, D. (2013). Alternative democratic assessment in PETE: an action-research study exploring risks, challenges and solutions. *Sport, Education and Society*, 18(1), 77-96. <https://doi.org/10.1080/13573322.2012.713859>
- McDermott, M. Z. R., Daniels, M., Nylén, A., Pears, A., Isomöttönen, V. y Caspersen, M. (2017). The authenticity of 'authentic' assessment some faculty perceptions. En M. E. Van Valkenburg (Ed.), *Proceedings of IEEE frontiers in education conference* (pp. 1-9). IEEE.
- Murphy, V., Fox, J., Freeman, S. y Hughes, N. (2017). "Keeping it Real": A review of the benefits, challenges and steps towards implementing authentic assessment. *AISHE-J: The All Ireland Journal of Teaching and Learning in Higher Education*, 9(3).
- Nguyen, H. (2017). *Authentic assessment in pedagogy-related modules in teacher education: Vietnamese student teachers' perspective*. University of East Anglia.
- Palacios, A. y López-Pastor, V. M. (2013). Haz lo que yo digo pero no lo que yo hago: sistemas de evaluación del alumnado en la formación inicial del profesorado. *Revista de Educación*, 361, 279-305. <https://doi.org/10.4438/1988-592X-RE-2011-361-143>.
- Pérez, A., Taberbero B., López Pastor, V. M., Ureña, N., Ruiz Lara, E., Caplloch, M., González Fernández, N. y Castejón, F. J. (2008). Evaluación formativa y compartida en la docencia universitaria y el Espacio Europeo de Educación Superior: cuestiones clave para su puesta en práctica. *Revista de Educación*, 347, 435-451.
- Prieto, J. P. A., Llácer, C. V. H. y Escobar, A. H. (2017). The effect of portfolios on higher Education students learning. En L. Gómez Chova, A. López Martínez, I. Candel Torres (Eds.), *Proceedings of 11th international technology, education and development conference* (pp. 6478-6480). IATED Academy.
- Quesada, V., Rodríguez, G. e Ibarra, M. S. (2013). ActEval: un instrumento para el análisis y la reflexión sobre la actividad evaluadora del profesorado universitario. *Revista de Educación*, 362, 69-104. <https://doi.org/10.4438/1988-592X-RE-2011-362-153>

- Rodríguez, G., Quesada, V. e Ibarra, M. (2016). Learning-oriented e-assessment: the effects of a training and guidance programme on lecturers' perceptions. *Assessment & Evaluation in Higher Education*, 41(1), 35-52. <https://doi.org/10.1080/02602938.2014.979132>
- Rule, A. C. (2006). The components of authentic learning. *The Journal of Authentic Learning*, 3(1), 1-10.
- Santos, J. M. (2017). Design, implementation and evaluation of an authentic assessment experience in a pharmacy course: are students getting it? En J. Domenech, M. C. Vincent Vela, E. Poza Plaza y M. D. Blázquez Soriano (Eds), *Proceedings of the 3rd international conference on higher education advances* (pp. 574-583). Universitat Politècnica de Valencia. <https://doi.org/10.4995/HEAD17.2017.5294>
- Sullivan, D. F. y McConnell, K. D. (2017). Big progress in authentic assessment, but by itself not enough. *Change: The Magazine of Higher Learning*, 49(1), 14-25. <https://doi.org/10.4995/HEAD17.2017.5294>
- Vu, T. T. (2011). *Investigating authentic assessment for student learning in higher education*. University of Queensland.
- Zabalza, M. Á. (2003). *Competencias docentes del profesorado universitario. Calidad y desarrollo profesional*. Narcea.

## Breve Cv de los autores

### **Emilio José Barrientos-Hernán**

Doctor en Educación y Licenciado en Psicopedagogía. Maestro de Educación Primaria especialidad en Educación Física, con 10 años de experiencia en Educación Primaria. ORCID ID: <https://orcid.org/0000-0002-6060-8486>. Email: [barrientosemilio@gmail.com](mailto:barrientosemilio@gmail.com)

### **Víctor M. López-Pastor**

Profesor de la Facultad de Educación de Segovia (Universidad de Valladolid). Catedrático de Universidad. Ha publicado 30 libros y numerosos artículos científicos y profesionales. Sus principales líneas de investigación son: \*Evaluación Formativa en educación superior, \*Evaluación Formativa en educación física, \*Formación del Profesorado, \*Investigación-Acción. ORCID ID: <https://orcid.org/0000-0003-2681-9543>. Email: [vlopez@mpc.uva.es](mailto:vlopez@mpc.uva.es)

### **Dario Pérez-Brunicardi**

Doctor en Educación (UVa) y Licenciado en Educación Física (UMP). Profesor Contratado Doctor de la Facultad de Educación de Segovia. Coordinador de la Red Estatal de Educación Física en la Naturaleza. Líneas de Investigación: Educación Física en la Naturaleza, Expresión Corporal, Deporte Escolar y Evaluación para el Aprendizaje. ORCID ID: <https://orcid.org/0000-0002-1347-1333>. Email: [dario.perez.brunicardi@gmail.com](mailto:dario.perez.brunicardi@gmail.com)



# La Evaluación Comprensiva de Programas Educativos: ¿Un Nuevo Paradigma Teórico?

## Comprehensive Evaluation of Educational Programs: A New Theoretical Paradigm?

Rafael López-Meseguer <sup>1\*</sup>

Manuel T. Valdés <sup>2</sup>

<sup>1</sup> Universidad Autónoma de Madrid y Fundación Europea Sociedad y Educación, España

<sup>2</sup> Universidad Complutense de Madrid, España

En el mismo año en que se cumplen cien años del fallecimiento de Max Weber, este trabajo explora la posibilidad de trasladar los principios de la “sociología comprensiva” a la evaluación de programas educativos. Para ello, en primer lugar y tras una revisión de la literatura sobre evaluación de programas educativos, se establece la relación entre el enfoque comprensivo de evaluación y otros enfoques evaluadores. En la segunda parte, se desarrolla la aportación teórica de este trabajo en términos de construcción del objeto de investigación-evaluación. En particular, se hace referencia al entendimiento de los programas educativos como relaciones sociales y a la evaluación de programas a partir del análisis de su coherencia lógica y su correspondencia causal. En la tercera parte, se definen los límites en cuanto a la realización de juicios de valor en la práctica evaluativa. Sobre este particular, en consonancia con Weber, se señala que, si bien las evaluaciones no deben realizar juicios de valor, ello no significa que no se puedan evaluar los juicios de valor de los participantes de los programas. En ese sentido, los juicios de valor de los participantes constituyen un elemento central dentro enfoque evaluativo que se presenta. Por último, se propone la articulación metodológica como la estrategia de evaluación más adecuada para dar cumplimiento a las prerrogativas teóricas de la evaluación comprensiva.

**Palabras clave:** Evaluación; Evaluación comprensiva; Sociología comprensiva; Programas educativos; Articulación metodológica.

In the 100th anniversary of Max Weber's death, this paper explores the possibility of translating the principles of "Comprehensive Sociology" to the evaluation of educational programs. To this end, we firstly review the literature on the evaluation of educational programs and establish the relationship between the comprehensive evaluation approach and other approaches. In the second part, we develop the theoretical contribution of this work in terms of the construction of the evaluation object. In particular, special attention is dedicated to the understanding of educational programmes as social relations and the evaluation of programs based on the analysis of their logical coherence and its causal correspondence. In the third part, we define the limits to value judgments in the evaluation practice. In line with Weber, we argue that, while evaluators should not make value judgements, it does not mean that participants' value judgements cannot be evaluated. In fact, they are a central element in the evaluation approach that we present. Finally, the articulation of qualitative and quantitative techniques is proposed as the most appropriate methodological strategy given the theoretical prerogatives of comprehensive evaluation.

**Keywords:** Evaluation; Comprehensive evaluation; Comprehensive sociology; Educational programs; Methodological articulation.

---

\*Contacto: rafaelmeseguer@gmail.com

## 1. Introducción

Llevar a cabo una evaluación implica decidir sobre el procedimiento a utilizar entre una amplia gama de posibilidades, y de ahí se deriva, naturalmente, la labor de definir adecuadamente qué constituye para cada cual el oficio de evaluar y la relación con aquello que se evalúa. La tendencia general de los evaluadores de programas, sin embargo, ha sido la de optar por un enfoque pragmático: en lugar de tratar de entender las relaciones entre las variables y su relación con la evaluación de programas en general (o del programa evaluado en particular), los teóricos de la evaluación se han centrado casi en exclusiva en facilitar la tarea de la evaluación a través de nuevos enfoques y diseños para mejorar la práctica evaluativa (Scriven, 1998).

Esta peculiaridad ha dado lugar a un inmenso cuerpo teórico sobre modelos evaluativos. Stufflebeam y Coryn (2014), en su célebre trabajo *Evaluation, Theory, Models & Applications*, son capaces de distinguir entre 23 modelos diferentes de evaluación respaldados con ejemplos particulares de evaluaciones llevadas a cabo. Lo común de estos casos es que todos ellos optan por unos u otros criterios epistemológicos y metodológicos particulares a la hora de llevar a cabo la práctica evaluativa. Conceptos como evaluación basada en criterios (Ligero, 2011), evaluación experimental y cuasi-experimental (Shadish et al., 2002), evaluación de estudio de caso (Stake, 1994, 1995), son habituales a la hora de distinguir maneras distintas de evaluar. Lo que se desprende a menudo de esta clase de distinciones es una escasa precisión terminológica y conceptual ya que, muchas veces, esas distinciones no diferencian correctamente lo que sería una técnica de evaluación, un modelo de evaluación y una teoría de la evaluación.

Una técnica de evaluación (la entrevista, la encuesta) sería la aplicación particular de una herramienta de ciencia social a la práctica de la evaluación de programas. Por su parte, entendemos que un modelo de evaluación es una concepción idealizada sobre cómo ha de llevarse a cabo la práctica evaluativa de la que se derivan unas prácticas más o menos estandarizadas. Una teoría de la evaluación, sin embargo, es algo más exigente; como señalan Stufflebeam y Coryn (2014, p. 50), es un conjunto coherente de principios conceptuales, analíticos y éticos que conforman un marco general para guiar el estudio y la práctica de la evaluación de programas. Se aprecia, por tanto, que la principal diferencia entre un modelo y una teoría radica en el grado de profundización sobre el objeto y la práctica de evaluar del que dispone el evaluador. Sin embargo, lo que hace a una teoría verdaderamente reconocible es que se sirve de unos conceptos particulares que informan sobre los fundamentos y el modo de conducir las evaluaciones en general, es decir, con los que comprender la totalidad del proceso evaluativo.

Por otro lado, hay que señalar que la teoría de la evaluación ha tenido un papel residual desde la emergencia de la evaluación como práctica profesional y académica reconocida. En un primer momento, las teorías iban dirigidas a contrarrestar algunas de las filosofías dominantes del gremio. Es así como se empezó a hablar de evaluación interpretativa (Stake, 1976) frente a la evaluación basada en criterios; de evaluación libre de objetivos (Scriven, 1973) frente a la evaluación estandarizada; de evaluación basada en la contingencia (Cronbach, 1982) o evaluación naturalista (Guba, 1978), frente a enfoques positivistas, etc. Sin embargo, los desarrollos teóricos de la evaluación de programas se fueron haciendo progresivamente más sofisticados y, en lugar de dedicarse a la crítica de la evaluación, comenzaron a desarrollarse teorías de la evaluación de carácter complejo

como puede ser el enfoque de la evaluación responsiva (Stake 2006), la evaluación CIPP (Stufflebeam, 2003), o la evaluación participativa (Cousins y Whitmore, 1998).

Llegados a este punto, es lícito preguntarse: ¿qué diferencia existiría, en la práctica, entre una teoría y un modelo? ¿Se derivan efectos específicos de la distinción entre una y otra? Si así fuera, ¿Qué habría de expresar una teoría para ser una teoría y no un modelo? En resumidas cuentas: ¿cómo podemos evaluar una teoría de la evaluación, por paradójico que suene?

A estas preguntas tratan de responder, en parte, William Shadish, Thomas Cook y Laura Levinton (1991) en su obra *Foundations of Program Evaluation: Theories of Practice*. En ella se señalan una serie de criterios con los que validar una teoría de la evaluación: (I) programación social, (II) construcción del conocimiento, (III) valoración, (IV) uso del conocimiento, y (V) práctica evaluadora. Siguiendo –en gran parte– lo establecido por los autores, consideramos que una teoría fundamentada de la evaluación debería responder, al menos, a: (I) la novedad de la teoría dentro de la tradición académica correspondiente, (II) cómo se construye el objeto de evaluación, (III) cuál ha de ser el tratamiento de los juicios de valor en las evaluaciones, y (IV) qué metodología(s) emplea en la práctica. Estos criterios deberían servir tanto para la evaluación de teorías de la evaluación como para la construcción de nuevas teorías. Así pues, en la medida que cada uno de estos interrogantes sean respondidos de una forma diferente a cómo se han respondido por parte de otras tradiciones teóricas, la pregunta que surge espontáneamente es la de si estaríamos frente a un nuevo paradigma teórico en el ámbito de la evaluación de programas en general, y de programas educativos en particular.

No obstante, con carácter previo al desarrollo de cómo el enfoque teórico de evaluación comprensiva que proponemos da respuesta a cada uno de los elementos anteriormente mencionados, consideramos necesario hacer una precisión acerca de la terminología empleada. En el contexto iberoamericano, es muy posible que el concepto de evaluación comprensiva pueda ser conocido en el ámbito académico. Ello se debe, sin embargo, a un criterio más de carácter lingüístico que evaluativo: y es que quién tradujo la obra de Robert Stake (2006), *Standards-based and responsive evaluation*, lo hizo como “evaluación comprensiva y evaluación basada en estándares”. Lo cierto es que el concepto “responsivo” o “respondiente” no existe en nuestro idioma, y en el momento de la traducción de la obra su utilización no era muy habitual por los científicos sociales ni por los evaluadores de programas. De ahí que, en los principales manuales de evaluación de programas en castellano, Stake haya quedado reflejado como el principal exponente de la evaluación comprensiva, cuando realmente debiera ser conocido como el exponente de la evaluación responsiva.

En ese sentido, conviene empezar señalando que nuestra utilización del concepto de comprensivo se basa en el intento de responder a los interrogantes teóricos mencionados desde las posibilidades que ofrece la “sociología comprensiva” de Max Weber (2014), en la que claramente se inspira, sumado a algunas aportaciones de otros importantes teóricos sociales (entre los que destacan Bourdieu, Chamboredon y Passeron) y de la evaluación (entre los que se incluye a Robert Stake, pero también muchos otros). Además, en el mismo año en que se conmemora el centenario del fallecimiento de Weber, este trabajo aspira a ser un ejemplo más de la actualidad del autor, con razón uno de los científicos sociales más afamados.

## 2. Teoría de la evaluación de programas

Existen numerosas formas de establecer distinciones teóricas en el ámbito de la evaluación de programas, aunque los criterios de esas distinciones varían de un autor a otro (Chen, 1990; Stake, 2006; Stufflebeam y Coryn, 2014). Como señala Stake (2006, p. 11), la primera labor que ha de realizar quien se adentra en el terreno de la teoría de la evaluación es exponer su propia clasificación de enfoques evaluadores. Este autor, por ejemplo, distingue entre evaluaciones “basadas en criterios y estándares” y evaluaciones “experienciales” o “interpretativas”. Otros autores, más interesados en disputas metodológicas, distinguen entre evaluación cuantitativa y cualitativa. Estos dos casos responderían a diferenciaciones epistemológicas y metodológicas.

Ramón Pérez Juste (2017), por su parte, distingue entre evaluaciones de procesos, evaluaciones de resultados y un último tipo de evaluaciones que califica como complejas. La evaluación comprensiva de programas educativos, como se verá más adelante, se circunscribe dentro de esta última categoría, por lo que el criterio clasificatorio que emplea este autor nos resulta más conveniente a efectos expositivos. En este caso, la diferenciación se establece a partir de un criterio teórico de base (el propio objeto de evaluación) y la distinción, por tanto, radica en el énfasis que se hace sobre uno u otro elemento del programa a la hora de valorarlo. Conviene señalar, no obstante, que, como todo criterio clasificatorio, se trataría de un artificio teórico con el que ordenar la realidad, sin que esta se corresponda del todo con dicho criterio.

Siguiendo con el argumento anterior, podríamos decir que los modelos de evaluación de resultados se dirigirían principalmente a corroborar si los resultados esperados ocurrieron y en qué medida son imputables al programa. Bajo esta rúbrica podemos ubicar, en primer lugar y por orden de importancia, los estudios causales experimentales y cuasi-experimentales y, segundo, las propuestas evaluativas que siguen estándares o criterios preordenados. Los modelos de evaluación de procesos, por su parte, tratarían de comprender los mecanismos en juego a la hora de vincular los procesos y resultados de una determinada intervención. Aquí se dan cita diferentes conceptos como los de “teoría del programa” (Weiss, 1972), “evaluación basada en la teoría” (Chen, 1990; Fitz-Gibbon y Morris, 1975) y, más recientemente, “teoría del cambio” (Rogers, 2014), que se diferencian entre sí por enfatizar o priorizar algunos elementos del proceso sobre otros.

### 2.1. *Enfoques teóricos de evaluación de resultados*

Dentro de los enfoques de resultados, los diseños experimentales son los que gozan de mayor popularidad. Su aplicación está dirigida a la atribución causal de los resultados observados al programa, esto es, tienen por objetivo analizar la relación de causalidad entre el cambio observado en una variable resultado y la aplicación de un programa de intervención (Stufflebeam y Coryn, 2014). Para ello, se construyen dos grupos de individuos, uno que pasará por el programa evaluado y que habitualmente se denomina “grupo de tratamiento”, y otro que no pasará por el programa, ya sea por pasar por un programa alternativo o por no pasar por ningún programa, que constituye el “grupo de control”. A fin de poder llevar a cabo esa atribución causal, la asignación de casos a los grupos de control y tratamiento debe ser aleatoria (Shadish, Cook y Campbell, 2002), garantizando así una distribución equilibrada en ambos grupos de posibles variables perturbadoras; es decir, permitiendo al evaluador descartar todas las explicaciones alternativas de los resultados observados excepto la propia aplicación del programa. Dicha aproximación es a menudo conocida como evaluación contrafactual, debido a que el grupo

de control funcionaría como un escenario donde observamos qué hubiese ocurrido con los casos tratados de no haber pasado por el programa (Imbens y Rubin, 2015).

No obstante, su implementación no es siempre posible, en tanto que requiere considerables recursos, un elevado compromiso de los implementadores del programa y un alto grado de colaboración por parte de los sujetos participantes en la evaluación (Stufflebeam y Coryn, 2014). Es por eso que es habitual recurrir a evaluaciones cuasi-experimentales, donde se relaja el supuesto de aleatorización en la asignación de los casos a los grupos de control y tratamiento. Sin embargo, al no haber distribuido los casos al azar, es posible que ciertos individuos de ciertas características se concentren en el grupo de tratamiento y, a la vez, que sean esos mismos individuos quienes manifiesten una mejoría en los resultados analizados. Dicha situación podría confundir al evaluador y hacerle pensar que es el programa quien causa los resultados cuando, en realidad, son esas ciertas características de los individuos del grupo de tratamiento las responsables del cambio observado (Imbens y Rubin, 2015). En tales circunstancias es necesario tomar medidas para poder atribuir causalmente los resultados al programa o descartar que éste tenga algún impacto, siendo habitual recurrir a técnicas de emparejamiento de casos tratados y de control, a diseños de regresión discontinua o a diseños de series temporales interrumpidas, entre otros procedimientos (Stufflebeam y Coryn, 2014).

Por último, es interesante destacar los modelos de evaluación por criterios, evaluación preordenada o evaluación basada en estándares, donde la valoración de un programa se hace en base a ciertos criterios y estándares preestablecidos (Ligero, 2011, p. 5), incrementando así la comparabilidad de los resultados alcanzados entre distintos equipos evaluadores, contextos geográficos, poblaciones sometidas al programa, etc. Este tipo de enfoques emergieron con fuerza durante los años sesenta en el ámbito de la cooperación al desarrollo y, con el paso del tiempo, fueron trasladándose hacia otros ámbitos. En el campo educativo, la creciente disponibilidad de bases de datos internacionales estandarizadas ha provocado una atención casi exclusiva hacia estos enfoques evaluativos, ya que permiten tener referencias nacionales, internacionales y regionales desde una perspectiva sistémica. Son destacables en ese sentido los esfuerzos realizados por la OCDE impulsando las pruebas PISA (*Program for International Student Assessment*) y PIACC (*Program for International Assessment of Adult Competences*), y por la IEA (*International Association for the Evaluation of Educational Achievement*) a través de las pruebas PIRLS (*Progress in International Reading Literacy Study*) y TIMMS (*Trends in International Mathematics and Science Study*).

Pese a que tales esfuerzos han supuesto una oportunidad inigualable para disponer de información que facilite la toma de decisiones en materia de política educativa, no puede dejar de reconocerse que no son evaluaciones que informen, por ejemplo, sobre los resultados que estén teniendo distintas formas de innovación docente o formas alternativas de relación entre la escuela y la comunidad. De ahí que, junto con las evaluaciones externas del sistema educativo (nacionales e internacionales) y los modelos de evaluación interna (a nivel de centro, del profesorado y del alumnado), deba establecerse un espacio para la evaluación de programas, de modo que sea posible evaluar cómo responde la lógica de un programa de intervención a su aplicación en distintos contextos educativos.

## **2.2. Enfoques teóricos de evaluación de procesos**

Los modelos de evaluación de procesos surgen, en cierta medida, como reacción frente a la primacía de los modelos de evaluación de resultados, ya que tendrían serias dificultades para responder a interrogantes de suma importancia a la hora de valorar un programa como, por ejemplo: ¿a qué se debe la eficacia o ineficacia del programa? ¿Qué elementos del programa son los que hacen que funcione y cuáles no? ¿Qué lugar ocupa y qué importancia tiene la experiencia vivida por los participantes? De ahí que surjan una serie de propuestas evaluativas que ponen el foco en el programa en sí y en su proceso de implementación.

El concepto de teoría del programa, del que emanan todos los enfoques centrados en los procesos, se popularizó con el trabajo de Edward Suchmann (1967), *Evaluative Research*. En él se presentan dos tipos de razonamientos que podrían explicar el fracaso de un programa: (I) que el programa falle en el intento de poner en marcha las actividades (*implementation failure*), y (II) que el fallo se deba a que el programa no genera los resultados previstos, ya que lo que prescribe no tenga consistencia (*theory failure*). Las evaluaciones de resultados, por sí solas, no proporcionan información alguna sobre ninguno de los aspectos mencionados. Es por eso que, aunque de modo minoritario, en la década de los setenta comienzan a aparecer una serie de trabajos que, desde diferentes perspectivas, enfatizan la importancia de analizar lógicamente las teorías que emergen de los programas, y tratan de desplegar mecanismos de evaluación que ayuden a averiguar la relación entre las actividades y los resultados del programa (Fitz-Gibon y Morris, 1975; Chen y Rossi, 1980; Weiss, 1997).

Así, por ejemplo, Chen (1990) sostiene que la teoría detrás del programa debe tener sentido en el marco de la teoría social. Para ello, considera necesario distinguir entre teoría normativa y teoría causal del programa. La primera serviría al propósito de establecer los objetivos y resultados que deberían ser perseguidos y posteriormente evaluados, mientras que la segunda se referiría al conjunto de proposiciones acerca de cómo funciona el programa, y que sería lo que se persigue validar científicamente.

Carol Weiss (1997), por su parte, distingue dos aspectos de la evaluación de programas: la teoría de la implementación (*implementation theory*) y la teoría programática (*programmatic theory*). La primera se centra en cómo se ha llevado a cabo el programa, es decir, testar si el programa ha sido conducido según lo planeado para llegar a los resultados esperados. La teoría programática, por su parte, indaga sobre los mecanismos que intervienen entre el desarrollo del programa y la generación de unos resultados esperados. Lo importante, para esta autora, no es tanto las actividades per se, sino la respuesta generada a tales actividades. Y es precisamente esa atención a la respuesta de los participantes de los programas la que ha abierto la puerta a las técnicas cualitativas de evaluación hasta convertirse en una práctica habitual de las evaluaciones de procesos. Por otro lado, cabe señalar que Weiss ha acabado por emplear la terminología de “evaluaciones de teoría del cambio” (*theories of change evaluation*) para designar el proceder evaluativo que combina teoría de la implementación y programática. Este concepto, además, se ha convertido en el más popular entre las evaluaciones de procesos.

Patricia Rogers es una de las autoras que más ha desarrollado esta perspectiva. Para la autora, “la teoría del cambio explica cómo se entiende que las actividades produzcan una serie de resultados que contribuyen a lograr los impactos finales previstos (...). En ocasiones, el término denomina de manera genérica a cualquier versión del proceso; por

ejemplo, a una cadena de resultados con una serie de cuadros de insumos vinculados a productos, resultados e impactos, o a un marco lógico que expone la misma información en una matriz” (Rogers, 2014, p. 1).

En definitiva, en lo que coinciden los modelos de evaluación de procesos es en la necesidad de plantear un marco lógico con el que ser capaces de entender cómo un determinado programa produce cambios en la realidad, y cómo relacionar esos cambios con los posibles efectos que dicho programa pueda generar. Y para que esto sea posible resulta necesario detenerse en cómo esos cambios son asumidos por los participantes del programa. Además, es importante destacar la importancia de estos enfoques en hacer posible la generalización de los procesos de implementación y posterior escalabilidad de los programas. En la medida que se dispone de información sobre el funcionamiento del programa surge la posibilidad de, por un lado, mejorar aquellos aspectos que no hayan funcionado correctamente y, por el otro, sistematizar los procesos de implementación a partir de la información recabada, de manera que pueda ser llevado más fácilmente a la práctica en otros contextos.

### ***2.3. Enfoques teóricos de evaluación complejos***

La forma en la que interactúan procesos y resultados, sin embargo, no agota la cuestión de los enfoques de evaluación: puede ocurrir que programas con una lógica bien establecida y bien diseñados técnicamente fracasen, y que la respuesta a dicho fracaso esté fuera de la lógica de los procesos y de los resultados. De ahí que, a menudo, para llevar a cabo la práctica de evaluaciones se requieran enfoques más amplios que llamamos, siguiendo a Pérez Juste (2017), modelos de evaluación complejos.

Dentro de este tipo de evaluaciones podríamos incluir al mencionado Stake (2006) y su modelo de evaluación responsiva, el modelo CIPP de Stufflebeam (2003), o la propuesta evaluativa del propio Pérez Juste (2017). Sin entrar al detalle de cada una de estas propuestas evaluativas, se podría decir que Stake pone un gran énfasis en la apreciación valorativa, tratando de deslocalizarla y situarla lejos de la apreciación del evaluador, dando lugar a diferentes “jueces” que se encargarían de analizar la “congruencia” y “contingencia” del programa, siguiendo su propia terminología. En ese sentido, como se verá más adelante, la importancia otorgada a los juicios valorativos de los participantes por este autor comparte una de las premisas fundamentales del enfoque de evaluación comprensiva que desarrollamos en este trabajo.

Stufflebeam, por su parte, opta por un enfoque más pragmático: su modelo CIPP (contexto, input, proceso y producto) hace un intento ciertamente fructuoso de síntesis entre enfoques de resultados y de procesos, a lo que suma una atención particularizada y atenta al contexto en el que se desarrolla el programa.

Por último, Pérez Juste, con una propuesta de estructuración muy próxima a la de Stufflebeam, periodiza la evaluación en cuatro momentos: la evaluación del programa en cuanto tal, la evaluación del proceso de implantación del programa, la evaluación de los resultados del programa, y la institucionalización de la evaluación del programa. Este último elemento es particularmente interesante, ya que considera que cada evaluación debe servir para la propia mejora de la práctica evaluadora, y que esa reflexión debe ser explícita e incorporarse al modelo.

En definitiva, podríamos señalar que lo que caracteriza a estos modelos y les dota de un carácter complejo es que parten de paradigmas teóricos que les permiten integrar los

análisis empíricos sobre los programas dentro de enfoques interpretativos más amplios. Para que ello sea posible, resulta necesario explicitar la relación entre el evaluador y el objeto de evaluación; ofrecer una tesis sobre el tratamiento de los juicios de valor, es decir, definir los elementos a partir de los cuales se puede decir que un programa es bueno o malo, eficaz o ineficaz; y, por último, señalar la metodología que se ha de seguir para llevar lo anterior a término. A continuación, se describe la forma en que la evaluación comprensiva concibe cada uno de esos elementos.

### **3. Construcción del objeto de evaluación: los programas educativos como relaciones sociales**

En su célebre trabajo, *El oficio del sociólogo*, Bourdieu, Chamboredon y Passeron (2002, p. 50) afirman, en referencia a Saussure, que es el punto de vista el que crea el objeto. Esa definición es completada por Weber, quien sostenía que no son las relaciones reales entre cosas las que delimitan los saberes científicos, sino las relaciones conceptuales entre problemas.

Para adentrarse en el ámbito de la evaluación, el evaluador siempre debe operar con unos conceptos que le permitan romper con las formas habituales de entender el objeto de evaluación ya que, de lo contrario, estas formas de pensar espontáneas se introducirían en la perspectiva del evaluador de manera inconsciente. En este sentido, siguiendo a Durkheim, la ruptura consistiría en la elaboración de nociones científicas, es decir, la definición previa del objeto como construcción teórica “provisoria” destinada a “sustituir las nociones del sentido común por una primera noción científica” (Maus, *texto 5*; en Bourdieu, Chamboredon y Passeron, 2002).

Con los conceptos de ruptura (epistemológica) y construcción del objeto lo que se quiere poner de manifiesto es que todo procedimiento evaluativo necesita contar con una suerte de esquemas reflexivos previos para poder llevarse a cabo. Esos esquemas constituyen lo que Bourdieu, Chamboredon y Passeron denominan el “oficio del sociólogo”. Análogamente, nosotros hablamos de “el oficio del evaluador”. Pero ese oficio, sin embargo, no ha de interpretarse como si de una receta de cocina se tratase: se añaden los ingredientes, se llevan a cabo las elaboraciones en un sentido pautado y el producto es la evaluación. De lo que se trataría es de hacer explícita la visión del evaluador a la hora de aproximarse al objeto de evaluación. De ahí que, a nuestro juicio, una teoría de la evaluación deba ser entendida como un conjunto de esquemas reflexivos que atraviesan la relación entre el evaluador y el objeto de evaluación<sup>1</sup> y que, en el caso de la evaluación comprensiva, estarían basados en (I) el entendimiento de los programas educativos como relaciones sociales, y (II) la evaluación de los programas a partir del análisis de su coherencia lógica y su correspondencia causal.

Sobre el primero de los aspectos, Max Weber (2014, p. 117) llama relación social al “comportamiento de varias personas en la medida en que el significado de la acción de cada una esté referido al de las otras y la acción se guíe por esa referencia”. Si tomamos como referencia esta definición, en el ámbito educativo podemos encontrar muchos tipos de

---

<sup>1</sup> William Shadish (1998) sintetizó esta idea de la mejor manera al señalar que “la teoría de la evaluación es lo que somos” (*evaluation theory is who we are*).

relaciones sociales susceptibles de ser evaluadas como programas: desde un proyecto que busque promover las competencias sociales y cívicas en el alumnado de secundaria a través del aprendizaje por proyectos, a un curso de formación que busque mejorar las habilidades de dirección de los equipos directivos; desde programas de implantación de títulos universitarios que busquen integrar unas competencias transversales comunes, hasta programas que busquen mejorar el clima entre el profesorado del centro. Lo fundamental, en ese sentido, no es tanto la casuística que cabe incluir bajo la etiqueta de programas, que puede ser mucha y variada, sino el “sentido” que guíe la acción (mejorar la competencia social y cívica de los adolescentes, las habilidades directivas, las competencias transversales de los universitarios, mejorar el clima entre el profesorado). De la consideración de los programas educativos como relaciones sociales se sigue, por tanto, un principio de reciprocidad con respecto a los fines del programa que debe ser identificable por los participantes.

En ese sentido, para que exista una relación social las partes que participan de ella deben estar sujetas a un cierto reconocimiento mutuo, es decir, que en mayor o menor medida reconocen que la acción que realizan está guiada por la expectativa de lo que aquella relación significa. Llevado al ámbito que nos ocupa, un programa prescribe un conjunto de actividades a realizar guiados por un fin que expresa aquello que se espera lograr. De ahí que, para que un programa tenga sentido, ese sentido debe ser reconocido por los participantes del mismo. No obstante, el grado y modo particular en el que ese reconocimiento se produce será, naturalmente, cambiante: cuando ese acuerdo es total, algo que ocurre solo en la teoría, nos encontraríamos frente a un tipo ideal<sup>2</sup> “puro” de programa (es decir, que el programa funcionaría a la perfección en relación a los fines que se propone); cuando ese acuerdo sobre los fines, actividades y logros del programa es suficiente, habría que valorar el grado de correspondencia causal de dicho programa, esto es, el grado de dicha suficiencia; cuando esa correspondencia no existe, significaría que el programa carece de consistencia y que, por tanto, no tiene sentido que pueda producir efectos.

Para situar esta idea en la práctica de programas educativos, consideremos el ejemplo anteriormente mencionado de un programa que busca mejorar las habilidades de dirección de una serie de equipos directivos de escuelas públicas. Imaginemos que ese programa tuviera entre sus fines el intento de mejorar las perspectivas de los equipos directivos acerca de la importancia de elevar las expectativas (personales, académicas, laborales) del alumnado susceptible de abandono educativo, y se les ofrece una formación específica sobre ello. Para que dicho programa fuera eficaz, en primer lugar, los participantes de dicho programa (los equipos directivos), deberían ser capaces de reconocer los fines de aquello que se les desea transmitir (la importancia de elevar las expectativas para evitar el abandono). En la medida que ello no ocurra, no se producirán efectos atribuibles al programa. En la medida que los equipos directivos reconozcan la validez de la idea, habrá que valorar en qué medida y cómo esa idea transforma sus significados acerca de la

---

<sup>2</sup> Recordemos que, Para Weber, los tipos ideales no son leyes generales de las cuales el fenómeno individual sea un ejemplo, sino un concepto abstracto, “relativamente vacío respecto a la realidad concreta” (Weber, 2006: 104). En ese sentido, el tipo ideal representa el grado de racionalidad perfecta que cabría esperar de un programa educativo.

importancia de las expectativas del alumnado susceptible de abandono, y en qué medida y cómo esta idea orienta sus acciones directivas después de recibir la formación.

Por otro lado, de la consideración de los programas educativos como relaciones sociales se derivan una serie de consecuencias epistemológicas. En uno de sus escritos fundamentales de teoría sociológica, *La objetividad del conocimiento en la ciencia social y en la política social*, Weber argumentaba que la sociología debía tratarse de una ciencia de la realidad (Weber, 2009, en Abellán, 2014):

*Queremos comprender la realidad de la vida que nos rodea y en la que estamos inmersos en su peculiaridad, es decir, queremos comprender, por un lado, el contexto de sus fenómenos concretos en su forma actual y en su significación en la cultura, y, por otro lado, el motivo de que hayan sido así y no de otra manera. (p. 17)*

Weber apuesta por una explicación causal de los fenómenos sociales, pero esa causalidad no ha de buscarse en la subsunción de fenómenos bajo el paraguas de leyes generales, al modo de las ciencias naturales. La comprensión (*verstehen*) de los fenómenos sociales y culturales requiere captar su individualidad (Abellán, 2014), y captar la individualidad de los fenómenos requeriría atender al significado subjetivo atribuido por los sujetos a las acciones. La tradición sociológica ha llamado a este modo de proceder individualismo metodológico. Aunque Max Weber no se ocupó de aplicar su teoría al campo educativo, su pensamiento influyó de forma significativa en algunos paradigmas teóricos cuya influencia en este campo ha sido notable, entre los que se encuentra el interaccionismo simbólico. Para los interaccionistas los seres humanos actuarían respecto a las cosas de acuerdo con los significados que éstas tienen para ellos<sup>3</sup>.

En ese sentido, la reconstrucción de los puntos de vista subjetivos de los participantes del programa sería uno de los objetos centrales de la evaluación comprensiva de programas educativos. Volviendo sobre el ejemplo anterior, diríamos que para considerar si el programa ha tenido efectos en cuanto a la elevación de las expectativas del alumnado susceptible de abandono por parte de los equipos directivos, un elemento fundamental de enjuiciamiento sería la propia consideración de los equipos directivos acerca del efecto que ha tenido el programa sobre las expectativas que depositan en su alumnado tras la formación específica. Ello encuentra justificación en que todo ser humano construye ideas, pensamientos y juicios sobre y a partir de sus propias experiencias, y eso le lleva a concebir el mundo de forma diferente al resto, lo que, a su vez, tiene consecuencias en su modo de actuar (Murillo, Hidalgo y Flores, 2016).

La segunda premisa es que la evaluación comprensiva basa su acción en el análisis de la coherencia lógica y la correspondencia causal de los programas. La conexión de los motivos de las relaciones sociales (o programas) con este tipo de evaluación la describe el propio Weber (2014):

*Motivo es el conjunto de elementos que se le presenta al observador o al propio agente como fundamento que da el significado al comportamiento. Decimos que un comportamiento que se*

---

<sup>3</sup> Herbert Blumer, en su célebre trabajo, *Symbolic Interactionism. Perspective and Method*, resumió perfectamente el principal andamiaje teórico de una tradición por entonces emergente (1969, p. 2 en Flick, 2004, p. 32): “la primera premisa es que los seres humanos actúan con respecto a las cosas de acuerdo con los significados que éstas tienen para ellos. La segunda premisa es que el significado de estas cosas se deriva o surge de la interacción social que se tiene con los propios compañeros. La tercera premisa es que estas cosas se manejan en un acuerdo interpretativo utilizado por la persona al hacer frente a las cosas que encuentra, y este proceso las modifica”.

*realiza coherentemente tiene coherencia lógica en la medida de que, de acuerdo con los usos promedio de nuestro pensamiento y de nuestros sentimientos, podamos afirmar que la relación existente entre los distintos elementos del comportamiento constituye un motivo típico (solemos decir que es un motivo lógico o correcto). Decimos, por el contrario, que una sucesión de fenómenos tiene correspondencia causal en la medida en que -de acuerdo con las reglas de la experiencia- existe la probabilidad de que esa sucesión se dé siempre en la realidad de la misma manera. (p. 90)*

Los motivos que inspiran los comportamientos van de la mano de las acciones observables en la realidad social, y es precisamente esa conexión la que nos da una medida comprensiva del programa. De ahí que afirmamos que un programa es coherente lógicamente cuando (I) sea posible afirmar, en base al conocimiento científico existente y a las reglas de la experiencia común, que aquello que el programa prescribe tiene fundamento lógico y, (II) cuando los fines del programa hayan sido comúnmente reconocidos por los participantes en grado suficiente. Por otro lado, decimos que un programa tiene correspondencia causal cuando (I) se atestigua que los participantes del programa se comportan conforme a las prescripciones del mismo, es decir, que el programa produce cambios en los comportamientos o en la realidad educativa, y (II) cuando esos cambios se deben al programa objeto de evaluación.

En el ejemplo que estamos analizando, diríamos, por tanto, que el programa tendría fundamento lógico en la medida que aquello que prescribe se adecúa al conocimiento sobre la materia (elevar las expectativas de los alumnos es uno de los principios más mentados en la literatura sobre liderazgo educativo); que los fines sean reconocidos por los participantes en grado suficiente, es decir, que reconozcan la importancia de elevar las expectativas de los alumnos susceptibles de abandono, aunque cada uno lo haga de manera diferente en función de las concepciones previas sobre tales expectativas; y que, fruto de la formación (correspondencia causal), tomen medidas que se dirijan a elevar las expectativas de sus alumnos como, por ejemplo, invitar a antiguos alumnos que hayan tenido éxito en sus trayectorias académicas o profesionales, con el objetivo de que los alumnos puedan reconocerse en la experiencia de otros que, estando en situaciones similares, se abrieron en paso en la vida.

La perfecta adecuación entre los objetivos del programa, la comprensión de los mismos por los participantes, las acciones llevadas a cabo y los resultados alcanzados constituiría, por tanto, el tipo ideal perfecto de grado de funcionamiento del programa, siendo lo demás desviaciones del tipo ideal. La coherencia lógica, en ese sentido, estaría en sintonía con la lógica de los enfoques de la evaluación de procesos, en cuanto que buscaría averiguar la correcta adecuación de objetivos, prácticas y resultados desde la perspectiva del evaluador y de los participantes. La comprobación de la correspondencia causal, por su parte, estaría más próxima a los enfoques de evaluación de resultados, ya que buscaría analizar el grado en que la situación típica, es decir, lo que fija el programa, ocurre en la realidad, ya sea a través de la perspectiva de los participantes o de otro tipo de pruebas estandarizadas con las que medir la eficacia del programa. Ambos elementos -la comprobación de la coherencia lógica y el análisis de la correspondencia causal- deben concurrir a la hora de valorar un programa desde un enfoque comprensivo de evaluación. En palabras del propio Weber (2014):

*Si no hubiera coherencia lógica estaríamos simplemente ante una probabilidad estadística no susceptible de comprensión (o solo comprensible de manera imperfecta), y toda vez que una afirmación solo puede ser causalmente correcta en la medida en que se aporte la prueba de que existe una determinada probabilidad (de alguna manera calculable) de que la acción suele*

*adoptar en la realidad -por término medio o en el caso ideal- un desarrollo lógicamente coherente con una frecuencia determinada o de modo aproximado.* (p. 91)

#### 4. Los juicios de valor en la práctica evaluativa

Evaluar, al fin y al cabo, no deja de ser una práctica dirigida a valorar algo. De lo que se trataría, por tanto, es de poder afirmar si un programa es bueno o malo; eficaz o ineficaz; si produce o no resultados. Por su parte, las ciencias sociales, desde su configuración como ámbito científico particular, han buscado establecer juicios valorativos que no dependan de valores particulares. Sin embargo, como señala Stake y todo evaluador reconoce, el juicio valorativo y la percepción se entremezclan a la hora de llevar a cabo cualquier evaluación (Stake, 2006, p. 15). Así pues, si queremos establecer un procedimiento de valoración que sea coherente con lo visto hasta ahora y que cumpla con los estándares de calidad exigidos a cualquier disciplina social, debemos, en primer lugar, realizar una primera aproximación al significado de los juicios de valor en la práctica de la evaluación de programas; y, en segundo lugar, proponer una estrategia metodológica que nos permita dar respuesta a las prerrogativas teóricas planteadas a partir de esa primera aproximación. A continuación, señalamos la manera en que la evaluación comprensiva responde a ambas cuestiones.

La defensa de Max Weber de una economía y una sociología libre de juicios de valor ha sido recibida como una especie de asunción dogmática por parte de la tradición sociológica, y ha dado lugar a multitud de interpretaciones erróneas acerca del pensamiento del autor. Es por eso que, de manera sintética, trataremos de exponer las principales ideas de Weber sobre la posibilidad de realización de juicios de valor en las ciencias sociales con el objetivo de situarla en el marco de la evaluación comprensiva de programas educativos.

Conviene señalar, con carácter previo, que el enunciado weberiano de abstenerse de hacer juicios de valor ha de concebirse, en cierta medida, como un modo de practicar la ciencia social diferente de la que practicaban los economistas e historiadores de su época y, también en cierta medida, como una asunción de los principios de reflexión de la escuela neokantiana, centrada en la libertad individual, con la que Weber mantenía una relación estrecha (Pérez Díaz, 1980, p. 66).

La pretensión de Weber es la de diferenciar, como ámbitos separados, la esfera de la explicación de los hechos y la esfera de los juicios de valor, para lo que insiste en la necesidad de distinguir entre “juicios de valor” y “relación con los valores” culturales (*Wertbeziehung*) y, consecuentemente, entre “juicio de valor” y “análisis de los juicios de valor” (Abellán, 2010). La intención última del autor, tal y como señala uno de sus más autorizados comentaristas, es, por un lado, mostrar su rechazo a la aplicación de los procedimientos científicos-naturales a las ciencias sociales y, por otro lado, renunciar a los conceptos “esencialistas” de la tradición romántica (Abellán, 2010, p. 22). Frente a la ciencia natural, que busca refrendar leyes objetivas de carácter general, las ciencias sociales han de poner el foco en la individualidad, porque de lo contrario se eliminaría todo aquello que constituye la peculiaridad de lo social.

Así pues, la relación con los valores se referiría a los compromisos que adquirimos con diferentes modos de vivir en comunidad, pero cuya validez depende o, mejor dicho, se circunscribe, a un contexto social e histórico determinado. Por tanto, se trataría de un concepto empírico de “valor”, por el que se supone que los sujetos se adscriben a valores culturales promovidos por diversas instituciones (familia, derecho, religión, estado,

costumbres, economía...). El estudio de la relación de los individuos con los valores sería la tarea de la ciencia social, mientras que, por el contrario, la valoración de dichos bienes culturales correspondería a la filosofía, ya sea esta concebida de un modo historicista o de manera contemporánea.

Weber, por tanto, establece así una distinción entre los juicios de valor y el análisis de los juicios de valor, y señala que entre uno y otro tipo de conocimiento hay una ruptura insalvable. Sin embargo, eso no significa que los juicios de valor no puedan ser analizados científicamente, y de ahí que estableciese un procedimiento que buscaba prescindir de valoraciones en el estudio de los juicios de valor subjetivos (Abellán, 2010: 39). La primera cautela a la que hacía referencia era la del peligro de asignar un propósito a estos análisis, es decir, que el estudio no nos debe llevar necesariamente a un acuerdo entre los juicios en discusión, sino todo lo contrario: puede ocurrir, y de hecho es normal que ocurra, que de la investigación (evaluación) se evidencie la imposibilidad de llegar a un consenso entre los juicios de valor en juego. Retomando el ejemplo anterior, se podría dar el caso en que uno de los directores de escuela que están participando en el programa considerase que elevar las expectativas de los alumnos es una forma de engaño, les sitúa fuera de la realidad, y ello podría tener efectos perniciosos. Es decir, que podría ser contrario a los fines establecidos por el programa, sin que pudiera haber un punto medio de encuentro y desplegar algún tipo de efecto.

La segunda cautela que establecía tenía que ver con la necesaria reflexión acerca de lo que pueden hacer y no hacer las ciencias sociales y, por ende, las evaluaciones con carácter científico, con respecto a la relación con los valores. En palabras del propio Weber (WL 508; en Abellán, 2010):

*En el terreno de los juicios de valor políticos -y especialmente en los juicios de valor de carácter económico o en los de política social-, si se quisieran deducir directrices en relación con una acción valiosa, lo único que una disciplina empírica puede ofrecer con sus propios medios es lo siguiente: 1) los medios indispensables (para esa acción valiosa), 2) las consecuencias inevitables y 3) las consecuencias prácticas de la competencia así generada entre los múltiples juicios de valor posibles entre sí. (p. 40)*

La ciencia empírica no puede ofrecer determinados tipos de explicaciones ni ha de emitir juicios de valor, ya que eso pertenece al saber de otras ciencias y, en último término, al ámbito privado del individuo. Sin embargo, que el científico social deba abstenerse de realizar juicios de valor no significa que no pueda analizar juicios de valor<sup>4</sup>. De hecho, para Weber, el análisis de los juicios de valor es un elemento fundamental, sino el más importante, de las ciencias sociales.

Trasladándonos de nuevo al enfoque de evaluación comprensiva, la consideración de los enunciados anteriores traería consigo una serie de consecuencias prácticas en su proceder metodológico: la primera operación evaluativa consistiría en formular los objetivos (primeros y últimos) del programa en términos de medios-fines racionales. En segundo lugar, se trataría de deducir las consecuencias derivadas de los fines que establece el programa, en coherencia con la evidencia existente. En tercer lugar, habría que comprobar el grado de adecuación entre los fines que establece el programa y los motivos que aducen los participantes. De estos tres elementos se encargaría precisamente la comprobación de

---

<sup>4</sup> De manera demostrativa y, por tanto, no exhaustiva, Weber estableció un procedimiento metodológico con el que analizar juicios de valor sin realizar juicios de valor (Abellán, 2010, p. 41).

la coherencia lógica del programa. En cuarto lugar, se trataría de observar si (I) lo que prescribe el programa ocurre, ya sea a través de la perspectiva de los participantes o de otro tipo de pruebas, (II) si ocurre en grado de probabilidad aceptable conforme a los fines propuestos, (III) si se han de tener en cuenta otras consecuencias inesperadas de su aplicación, o (IV) concluir que no hay una adecuación entre lo que el programa prescribe y lo que realmente ocurre. Esta operación es lo que llamamos el análisis de la correspondencia causal del programa.

No obstante, a pesar de ese valor central de los juicios de valor de los participantes, conviene señalar que no es el único método posible para atestiguar la correspondencia causal de un programa: en el ámbito educativo, así como en otras disciplinas sociales, suele ser habitual servirse de pruebas estandarizadas de tipo competencial, diagnóstica, u otras realizadas *ad hoc*, con las que medir los progresos de los participantes, en función de las características del programa. En el ejemplo que venimos utilizando, junto con la consideración por parte de los equipos directivos acerca de la mejora en términos de las expectativas académicas y profesionales del alumnado que le ha proporcionado el programa, podríamos tratar de establecer una medida sobre las expectativas que tienen los equipos directivos sobre el alumnado con carácter previo a la formación, y volver a medir con posterioridad a la misma, de tal forma que obtendríamos una medida de la mejora a través de un mismo indicador considerado en dos momentos distintos. Para asentar con mayor precisión el argumento, consideremos otro ejemplo hipotético de los anteriormente mencionados, como podría ser un plan dirigido a promover la competencia social y cívica de adolescentes a través del aprendizaje por proyectos. Junto con un test de competencias cívicas y sociales realizado antes y después de la implementación del programa, una evaluación comprensiva consideraría fundamental los juicios de los participantes (en este caso profesores y alumnos) acerca de los efectos que produce el programa.

En resumen, alrededor de los juicios de valor de los participantes del programa pivotan una serie de estrategias metodológicas que contribuyen a tener una medida comprensiva del programa. No obstante, el enfoque de evaluación comprensiva otorga una posición central a los juicios de valor de los participantes, sin los cuales no podría hablarse de evaluación comprensiva. La dificultad, por tanto, estribaría en cómo combinar todas estas estrategias de valoración. En ese sentido, no cabe duda de que el procedimiento que enuncia Weber, si bien todavía nos resulta útil para la reflexión epistemológica, requiere de algunas actualizaciones desde el punto de vista metodológico, pues hay que tener en cuenta los numerosos avances técnicos que han experimentado las ciencias sociales en los últimos cien años. A continuación, proponemos una estrategia metodológica con la que dar respuesta a los diferentes planteamientos teóricos expuestos hasta ahora.

## **5. Articulación metodológica de técnicas de investigación-evaluación**

En términos metodológicos, la evaluación ha ido sirviéndose progresivamente de los avances técnicos de la investigación social, aunque adaptados a sus propios fines, que no siempre han de ser coincidentes. De ahí que la terminología empleada en los debates metodológicos en uno y otro campo sea muy parecida e incluso intercambiable. En lo relativo a la evaluación comprensiva, la perspectiva adoptada es la de la articulación metodológica, donde las necesidades de información son satisfechas a través de estrategias

metodológicas que combinan técnicas cualitativas y cuantitativas. En el plano académico, tales estrategias son habitualmente referidas bajo múltiples denominaciones: métodos mixtos de investigación (Burke, Onwuegbuzie, y Turner, 2007), investigación multimétodo (Serrano et al., 2009), complementariedad metodológica (Blanco y Pirela, 2016), integración metodológica (Ruiz, 2008), etc. No obstante, la idea de articulación metodológica que sostenemos tendría un sentido más amplio: a la referida combinación de técnicas cualitativas y cuantitativas se le une que, en todo caso, las técnicas empleadas se han de orientar al análisis de la coherencia lógica y la correspondencia causal del programa objeto de evaluación.

Como es sabido, la relación entre la investigación cualitativa y cuantitativa ha sido históricamente controvertida. No faltan hoy detractores al intento de articulación de ambas aproximaciones, considerando que ambas constituyen paradigmas de investigación y evaluación inconmensurables por estar asentados en supuestos metodológicos, epistemológicos y ontológicos irreconciliables (Kuhn, 2005). Sin un acuerdo mínimo sobre cómo producir conocimiento, qué es conocimiento y qué es cognoscible, no hay posibilidad de articular una estrategia metodológica coherente. Y dado que los paradigmas cualitativo y cuantitativo en ocasiones no ofrecen las mismas respuestas para esas preguntas, las posibilidades de articulación se ven dificultadas. Cuando desde estas posiciones se habla de integración metodológica, a lo que se hace referencia es al recurso a técnicas distintas dentro de un mismo paradigma. Es lo que Ruiz (2008) denomina integración metodológica intraparadigmática.

Del otro lado, es habitual encontrar autores que evitan por completo ese debate y tienden a pensar en las técnicas cuantitativas y cualitativas como meros instrumentos de recogida de información. No es necesario, por tanto, dedicar mayor tiempo a reflexionar sobre las dificultades que pueda entrañar el recurso a distintas aproximaciones metodológicas o sobre la mejor manera de articular tales técnicas. El debate se centra, por tanto, en el plano puramente técnico, donde las posibilidades de integración son plenas e independientes de las circunstancias que caracterizan a una evaluación particular. Este sería un caso extremo de lo que Ruiz (2008) califica como integración metodológica interparadigmática.

Entre ambas posiciones existe un extenso espacio intermedio donde caben múltiples posibilidades de articulación. La propuesta que aquí se realiza, y que es compartida por otros muchos evaluadores, parte de los dos planteamientos siguientes: primero, que las aproximaciones cualitativa y cuantitativa presentan especificidades que requieren de una profunda reflexión acerca de las posibilidades reales de articulación en la evaluación particular que se tenga entre manos; y, segundo, que los objetivos y preguntas de evaluación específicos de cada caso no abocan irremediabilmente a un determinado paradigma, de manera que sea incoherente recurrir a metodologías propias del paradigma alternativo. Las distintas técnicas cualitativas y cuantitativas, por idiosincráticas que sean, podrán ser empleadas de forma conjunta si el evaluador reflexiona lo suficiente sobre cómo pueden ser mejor satisfechos sus objetivos de evaluación.

Al respecto de las posibilidades de integración o articulación metodológica, Blanco y Pirela (2016) distinguen tres estrategias: la combinación, la complementación y la triangulación. La combinación metodológica pretende una validación de resultados a través del recurso a técnicas de investigación cualitativas y cuantitativas que, implementadas de forma independiente, pueden llevar a resultados convergentes que refuercen las conclusiones alcanzadas. En ese sentido, el análisis de la correspondencia causal en el ejemplo sobre el liderazgo de los equipos directos podría valerse de caminos

separados que empleen, respectivamente, técnicas cualitativas (grupos de discusión con los equipos de dirección) y cuantitativas (mediciones previas y posteriores a la formación sobre la labor de dirección) con el objetivo de que converjan finalmente en unas conclusiones coherentes sobre las posibilidades de atribución causal de los resultados.

En segundo lugar, la complementación se basa en la elección de una perspectiva metodológica dominante, empleando la aproximación alternativa como una herramienta de profundización, matización o generalización. Es lo que se ha denominado como complementariedad asimétrica (Serrano et al., 2009), defendida por autores como Ibáñez (1986), o complementariedad por deficiencia (Ortí, 1995), enfatizando el conocimiento imperfecto de la realidad a que conducen por separado las aproximaciones cualitativa y cuantitativa (Martínez, 2005). En efecto, las técnicas de investigación cuantitativa podrían ser enormemente eficaces para estudiar la atribución causal de resultados a través de complejos procedimientos estadísticos. En el ejemplo sobre la formación en competencia cívica, podíamos llevar a cabo un diseño experimental con que tratar de atribuir la hipotética mejora al programa implementado. Pero dichos resultados solo serían inteligibles en el sentido de la evaluación comprensiva a través de las apreciaciones que permite el trabajo cualitativo, por ejemplo, con entrevistas a alumnos y profesores con los que indagar en el significado del concepto de competencia cívica y sobre si esta es o no enseñable.

En cuanto a la triangulación metodológica, el concepto busca acentuar la necesidad de complementar los resultados alcanzados por una vía a través de formas alternativas de investigación, donde sean empleados distintos evaluadores, perspectivas teóricas, fuentes de información y metodologías de análisis (Denzin, 1970; Denzin y Lincoln, 1994; Mertens y Hesse-Biber, 2012). Nótese que, si la combinación pretendía reforzar los resultados alcanzados llegando a las mismas conclusiones a través de aproximaciones metodológicas distintas, y la complementación buscaba emplear una de ambas aproximaciones como una herramienta auxiliar de la otra, la triangulación aspira a un esfuerzo de integración total y simétrico que ayude a superar reduccionismos y privilegios metodológicos. Por ejemplo, el análisis de la coherencia lógica podría valerse de un trabajo cualitativo exhaustivo que haga emerger los significados de los distintos participantes sobre la lógica del programa. Ese trabajo podría continuar después realizando un cuestionario que permita definir una tipología de acuerdo con variables como el sexo de los participantes, su edad y su rol en el programa.

Finalmente, consideramos muy relevante no ser excluyente en el análisis de la atribución causal de resultados y poner en valor lo que las técnicas cuantitativas y cualitativas pueden ofrecer en ese sentido. Sobre este particular, Howe (2012) distingue elocuentemente entre el estudio de las relaciones de causalidad mecánica y de causalidad agencial. Las primeras serían relaciones causales del tipo de las que ocurren en el mundo natural, donde A causa B de la misma manera en que la temperatura causa la evaporación. Este tipo de relaciones se encuentran siempre presentes en los programas de intervención y son aquellas típicamente estudiadas a través de evaluaciones contrafactuales. Pero esas no son todas las relaciones causales que conforman la lógica de un programa. También hay relaciones de causalidad agencial, basadas en el comportamiento intencional de actores, no en relaciones mecánicas. Desentrañar esas relaciones comporta un tipo de proceder evaluador diferente, donde las técnicas cualitativas pueden jugar un papel más relevante que las cuantitativas.

La conclusión última de todo ello es que, pese a que determinadas formas de evaluación han tendido a reducir el estudio de la causalidad a ciertas relaciones mecánicas, no existen barreras efectivas para que la articulación metodológica aborde, en el marco de una misma evaluación, relaciones causales mecánicas y agenciales a través de técnicas cualitativas y cuantitativas. De la misma forma se posiciona la propia Rogers (2014) al describir la teoría del cambio, para quien ésta “se sirve de un conjunto de datos cualitativos y cuantitativos para respaldar la triangulación de los datos obtenidos a raíz de una evaluación de métodos mixtos” (p. 7). Y es que desentrañar y comprobar la lógica de un programa y la correspondencia causal entre sus distintos elementos, en ocasiones implícitas para sus propios desarrolladores, no puede reducirse a la aplicación de una técnica de investigación particular, por poderosa que ésta pueda ser desde el punto de vista estadístico. Es necesario el desarrollo de una estrategia coherente que permita entender qué se esperaba del programa, cómo se esperaba que éste funcionase, qué ha ocurrido en la práctica y cómo se relacionan los resultados con su implementación real. La articulación metodológica sirve a dicho propósito.

## **6. Conclusiones**

En este trabajo hemos abordado las posibilidades que nos ofrece la sociología comprensiva de Max Weber en el ámbito de la evaluación de programas. Para ello, hemos empezado distinguiendo entre una teoría y un modelo de evaluación, hasta el punto de proporcionar nuestra propia definición de lo que es una teoría en relación a la evaluación de programas, considerando como tal un conjunto de esquemas y principios reflexivos que atraviesan la relación que se establece entre el evaluador y el objeto de evaluación.

Sobre los diferentes enfoques teóricos en el ámbito de programas, hemos distinguido entre enfoques de evaluación de resultados, de procesos y enfoques de evaluación complejos, donde hemos encuadrado el modo de evaluación comprensiva de programas educativos. Un enfoque de evaluación complejo sería aquel que fuese capaz de (I) proporcionar una teoría particular acerca de la relación entre el evaluador y el objeto de evaluación, (II) ofrecer una tesis sobre el tratamiento de los juicios de valor en los procesos de evaluación, es decir, definir los elementos a partir de los cuales se puede afirmar que un programa es bueno o malo, eficaz o ineficaz; y, por último, (III) señalar la metodología que se ha de seguir para llevar lo anterior a término. A lo largo del texto hemos ido describiendo el modo en que la evaluación comprensiva da respuesta a cada uno de estos interrogantes.

A propósito de la relación entre evaluador y objeto de evaluación, hemos señalado que la evaluación comprensiva se orientaría a partir de una serie de premisas o prerrogativas: (I) el entendimiento de los programas educativos como tipos ideales de relaciones sociales, y (II) la evaluación de los programas a partir del análisis de su coherencia lógica y su correspondencia causal. La primera de las premisas nos ha llevado a reconocer la importancia de la reconstrucción de los puntos de vista subjetivos de los participantes del programa como uno de los objetos centrales de la evaluación comprensiva, de tal forma que pudiera valorarse el grado de reciprocidad en cuanto a los fines que dicho programa pudiera prescribir. De acuerdo con la segunda premisa, hemos señalado que un programa es coherente lógicamente cuando (I) sea posible afirmar, en base al conocimiento científico existente y a las reglas de la experiencia común, que aquello que el programa prescribe tiene fundamento lógico y, (II) los fines del programa hayan sido comúnmente reconocidos por los participantes en grado suficiente. Por otro lado, decimos que un programa tiene

correspondencia causal cuando (I) se atestigua que los participantes del programa se comportan conforme a las prescripciones del mismo, es decir, que el programa produce cambios en los comportamientos o en la realidad educativa, y (II) cuando esos cambios se deben al programa objeto de evaluación.

En referencia al alcance del término valoración en el ámbito de la evaluación de programas, hemos conectado esta problemática con el clásico debate sobre los juicios de valor en las ciencias sociales para terminar optando por una solución de tipo weberiano: que la ciencia social deba estar libre de juicios de valor no quiere decir que no pueda analizar los juicios de valor. A partir de esta máxima, hemos afirmado que los juicios de valor de los participantes del programa son un aspecto central de la evaluación comprensiva, que ha de ser combinado, en la medida de lo posible, con otro tipo de pruebas estandarizadas, como suele ser habitual en el ámbito educativo.

Para llevar a la práctica lo descrito anteriormente, hemos optado por un procedimiento de articulación metodológica de investigación-evaluación, ya que la cuestión del método no puede reducirse a la aplicación de una técnica de investigación particular, por poderosa que ésta pueda ser desde el punto de vista estadístico. Por el contrario, toda evaluación comprensiva de programas educativos debe combinar las técnicas cualitativas y cuantitativas más eficaces orientadas al análisis de la coherencia lógica y la correspondencia causal del programa objeto de evaluación.

Por último, quedaría abierta la discusión con la que titulábamos este trabajo: ¿constituye la evaluación comprensiva un nuevo paradigma teórico en la evaluación de programas en general y en la evaluación de programas educativos en particular? Nuestra pretensión ha sido la de construir una serie de esquemas reflexivos, epistemológicos y metodológicos, que ayuden a avanzar en el terreno de la teoría de la evaluación y, por tanto, la decisión última acerca de si esta propuesta teórica constituye un nuevo paradigma le corresponde tomarla a quienes se reúnen dentro del oficio de la evaluación de programas.

## Agradecimientos

Queremos agradecer a la Fundación Europea Sociedad y Educación su compromiso con la evaluación de programas educativos. En particular, las reflexiones conjuntas en el marco de la unidad de Observación y Evaluación (OyE), han inspirado muchas de las ideas que se presentan en este trabajo.

## Referencias

- Abellán, J. (2010). Estudio preliminar. En M. Weber (Ed.), *Por qué no se deben hacer juicios de valor en la sociología y en la economía* (pp. 11-55). Alianza Editorial.
- Abellán, J. (2014). Estudio preliminar. En M. Weber (Ed.), *Conceptos sociológicos fundamentales* (pp. 11-71). Alianza Editorial.
- Blanco, N. y Pirela, J. (2016). La complementariedad metodológica: Estrategia de integración de enfoques en la investigación social. *Espacios Públicos*, 19(45), 97-111.
- Blumer, H. (1969). *Symbolic interactionism. Perspective and method*. University of California Press.
- Bourdieu, P. (2002). *El oficio de sociólogo: Presupuestos epistemológicos*. Siglo XXI Editores.

- Burke, R., Onwuegbuzie, A. J. y Turner, L. A. (2007). Toward a definition of mixed methods research. *Journal of Mixed Methods Research*, 1(2), 112-133.  
<https://doi.org/10.1177/1558689806298224>
- Chen, H.-T. (1990). *Theory-Driven Evaluation: A Comprehensive Perspective*. Sage.
- Chen, H.-T. y Rossi, P. (1980). The Multi-Goal, Theory-Driven Approach to Evaluation: A Model Linking Basic and Applied Social Science. *Social Forces*, 59(1), 106-122.  
<https://doi.org/10.2307/2577835>
- Cronbach, L. J. (1982). *Designing evaluations of educational and social programs*. Jossey-Bass.
- Cousins, J. B. y Whitmore, E. (1998). Framing participatory evaluation. En E. Whitmore (Ed.), *Understanding and practicing participatory evaluation* (pp. 5-23). Jossey-Bass.  
<https://doi.org/10.1002/ev.1114>
- Denzin, N. K. (1970). *Sociological methods: A sourcebook*. Watterworth.
- Denzin, N. K. y Lincoln, Y. S. (1994). *Handbook of qualitative research*. Sage.
- Fereday, J. y Muir-Cochrane, E. (2008). Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. *International Journal of Qualitative Methods*, 5(1), 80-92. <https://doi.org/10.1177/160940690600500107>
- Fitz-Gibbon, C. T. y Morris, L. L. (1975). Theory-based evaluation. *Evaluation Comment*, 5(1), 1-4.
- Flick, U. (2004) *Introducción a la investigación cualitativa*. Morata.
- Guba, E. G. (1978). *Toward a methodology of naturalistic inquiry in educational evaluation*. University of California Press.
- Howe, K. R. (2012). Mixed methods, triangulation, and causal explanation. *Journal of Mixed Methods Research*, 6(2), 89-96. <https://doi.org/10.1177/1558689812437187>
- Ibáñez, J. (1986). Cuantitativo/cualitativo. En R. Reyes (Ed.), *Terminología científico-social. Aproximación crítica* (pp. 218-233). Anthropos.
- Imbens, G. W. y Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press. <https://doi.org/10.1017/CBO9781139025751>
- Kuhn, T. S. (2005). *La estructura de las revoluciones científicas*. Fondo de Cultura Económica de España.
- Ligero, J. A. (2011). *Dos métodos de evaluación: criterios y teoría del programa. Documento de Trabajo, Serie CECOD N°15*. CEU Ediciones.
- Löffler, E. (1996). *La modernización del Sector Público desde una perspectiva comparativa. Conceptos y métodos para evaluar y premiar. La calidad en el sector público en los países de la OCDE*. Documentos INAP.
- Martínez, M. (2005). *El paradigma emergente: Hacia una nueva teoría de la racionalidad científica*. Trillas.
- Mertens, D. M. y Hesse-Biber, S. (2012). Triangulation and mixed methods research: Provocative positions. *Journal of Mixed Methods Research*, 6(2), 75-79.  
<https://doi.org/10.1177/1558689812437100>
- Murillo, F. J., Hidalgo, N. y Flores, S. (2016). Incidencia del contexto socio-económico en las concepciones docentes sobre evaluación. *Profesorado: Revista de Currículum y Formación del Profesorado*, 20(3), 251-281.
- Ortí, A. (1995). La confrontación de modelos y niveles epistemológicos en la génesis e historia de la investigación social. En J. M. Delgado y J. Gutiérrez (Eds.), *Métodos y técnicas cualitativas de investigación en ciencias sociales* (pp. 87-99). Síntesis.

- Pratt, D. D. (1992). Conceptions of teaching. *Adult Education Quarterly*, 42(4), 203-220. <https://doi.org/10.1177/074171369204200401>
- Pérez Juste, R. (2017). *Evaluación de programas educativos*. La Muralla.
- Pérez Díaz, V. (1980). *Introducción a la sociología: Concepto y método de la ciencia social en su historia*. Alianza Editorial.
- Remesal, A. (2011). Primary and secondary teachers' conceptions of assessment: A qualitative study. *Teaching and Teacher Education*, 27(2), 472-482. <https://doi.org/10.1016/j.tate.2010.09.017>
- Rogers, P. (2014). *La teoría del cambio*. Centro de Investigaciones Innocenti de Unicef.
- Ruiz, C. (2008). *El enfoque multimétodo en la investigación social y educativa: Una mirada desde el paradigma de la complejidad*. Upel.
- Scriven, M. (1998). Minimalist theory: The least theory that practice requires. *American Journal of Evaluation*, 19, 57-70. <https://doi.org/10.1177/109821409801900105>
- Serrano, A., Blanco, F., Ligeró, J. A., Alvira, F., Escobar, M. y Saén, A. (2009). *La investigación multimétodo*. Recuperado de [https://eprints.ucm.es/30034/1/araceli%20serrano%20articulacion\\_metodologica.\\_serrano\\_blanco\\_alvira.pdf](https://eprints.ucm.es/30034/1/araceli%20serrano%20articulacion_metodologica._serrano_blanco_alvira.pdf)
- Shadish, W. R. (1998). Evaluation theory is who we are. *American Journal of Evaluation*, 19, 1-19.
- Shadish, W. R., Cook, T. D. y Leviton, L. C. (1991). *Foundations of program evaluation: Theories of practice*. Sage. <https://doi.org/10.1177/109821409801900102>
- Shadish, W. R., Cook, T. D. y Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.
- Shavelson, R. J. y Towne, L. (2002). *Scientific research in education*. National Research Council.
- Suchman, E. (1967). *Evaluative research*. Russell Sage Foundation.
- Stake, R. E. (1976). A theoretical statement of responsive evaluation. *Studies in Educational Evaluation*, 2, 19-22. [https://doi.org/10.1016/0191-491X\(76\)90004-3](https://doi.org/10.1016/0191-491X(76)90004-3)
- Stake, R. E. (1994). Case studies. En N. K. Denzin y Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 236-247). Sage.
- Stake, R. E. (1995). *The art of case study research*. Sage.
- Stake, R. E. (2006). *Evaluación comprensiva y evaluación basada en estándares*. Graó.
- Stufflebeam, D. L. (2003). The CIPP model for evaluation. En T. Kellaghan y D. L. Stufflebeam (Eds.), *International handbook of educational evaluation* (pp. 31-62). Kluwer. [https://doi.org/10.1007/978-94-010-0309-4\\_4](https://doi.org/10.1007/978-94-010-0309-4_4)
- Stufflebeam, D. L. y Coryn, C. L. S. (2014). *Evaluation theory, models, and applications*. Jossey Bass.
- Thompson, A. G. (1992). Teachers' beliefs and conceptions: A synthesis of the research. En D.A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 127-146). Macmillan.
- Weber, M. (2009). *La "objetividad" del conocimiento en la ciencia social y en la política social*. Alianza Editorial.
- Weber, M. (2010). *Por qué no se deben hacer juicios de valor en la sociología y en la economía*. Alianza Editorial
- Weber, M. (2014). *Conceptos sociológicos fundamentales*. Alianza Editorial.

Weiss, C. H. (1972). *Evaluation research: Methods for assessing program effectiveness*. Prentice Hall.

Weiss, C.H. (1997). *Theory-Based Evaluation: Past, Present and Future*. Jossey-Bass Publishers.  
<https://doi.org/10.1002/ev.1086>

## **Breve CV de los autores**

### **Rafael López-Meseguer**

Investigador doctorando en la Facultad de Ciencias Políticas de la Universidad Autónoma de Madrid, donde obtuvo su graduado en Ciencias Políticas y de la Administración, y cursó el Master en Democracia y Gobierno. Investigador asociado al Instituto de Estudios Sociales y Educativos de la Fundación Europea Sociedad y Educación, donde colabora en diversos proyectos de investigación. Sus temas de especialización son la Teoría Política y Social, la Teoría de la Evaluación, y la Investigación cualitativa en Educación. ORCID ID: <https://orcid.org/0000-0001-8488-2888>. Email: [rafaelmeseguer@gmail.com](mailto:rafaelmeseguer@gmail.com)

### **Manuel T. Valdés**

Personal investigador en formación en el departamento de Sociología Aplicada de la Universidad Complutense de Madrid (ref: FPU16/02905). Egresado en Ciencias Políticas y Sociología por la Universidad Carlos III de Madrid y del Máster en Metodología de la Investigación en Ciencias Sociales de la Universidad Complutense de Madrid. Ha colaborado en investigaciones en el Centro Nacional de Medicina Tropical del Instituto de Salud Carlos III y en el Instituto Nacional de Estadística. Sus intereses de investigación se centran en la evaluación de políticas educativas, las desigualdades ante la educación y el papel de la educación en los procesos de movilidad social. Sus trabajos han sido publicados en revistas como la Revista Española de Investigaciones Sociológicas o la Revista Española de Sociología. ORCID ID: <https://orcid.org/0000-0001-8012-3956>. Email: [manueltv@ucm.es](mailto:manueltv@ucm.es).



## El Examen de Ingreso a la Universidad Nacional Autónoma de México: Evidencias de Validez de una Prueba de Alto Impacto y Gran Escala

### The Admission Exam to the National Autonomous University of Mexico: Validity Evidence of a Large Scale High-Stakes Test

Melchor Sánchez Mendiola \*  
Manuel García Minjares  
Adrián Martínez González  
Enrique Buzo Casanova

Universidad Nacional Autónoma de México, México.

**Introducción.** Los exámenes de admisión a la educación superior son evaluaciones sumativas de alto impacto para los aspirantes, por lo que requieren evidencia de validez para que las inferencias que se hagan de los resultados sean apropiadas. La Universidad Nacional Autónoma de México (UNAM) es la institución de educación superior más solicitada del país, anualmente ingresan menos del 10% de los aspirantes por examen de selección. **Métodos.** Se realizó un análisis de las fuentes de evidencia de validez del examen, con el modelo conceptual de Messick, Kane y los Estándares de la AERA-APA-NCME, con la información generada de la aplicación de febrero 2019 a 148.407 sustentantes. **Resultados:** Se identificaron evidencias de validez de contenido, proceso de respuesta, estructura interna, relación con otras variables y consecuencias del examen. Los resultados revelan que el examen de ingreso tiene suficiente evidencia de validez para afirmar que es sólido como herramienta de medición del conocimiento. **Discusión.** Por su relevancia social, es fundamental que las instituciones que usan este tipo de instrumentos documenten sus evidencias de validez. Es necesario realizar investigaciones periódicas longitudinales sobre el uso del examen, ya que las condiciones sociales y educativas del contexto de la población de aspirantes son dinámicas.

**Palabras clave:** Condiciones de admisión; Evaluación sumativa; Prueba de respuesta múltiple; Selección de estudiantes; Validez.

**Introduction.** Higher education institutions' admission exams are summative high-stakes tests that have important consequences for applicants, so they require validity evidence to assure that appropriate inferences are made with the results. The National Autonomous University of Mexico (UNAM) is the most sought-after higher education institution in the country, annually less than 10% of applicants that take the test are admitted. **Methods.** Analysis of the sources of the test validity evidence was performed using Messick and Kane conceptual frameworks, as well as the AERA-APA-NCME Standards, with the information generated from the February 2019 admission test in 148.407 applicants. **Results:** Test validity evidence was identified from content, response process, internal structure, relationship with other variables and consequences. Results suggest that the test has enough validity evidence, to state that the instrument is robust as a technical tool for knowledge assessment and as a source of information for high-stakes decisions. **Discussion.** It is crucial that institutions that use these tools document their validity evidence, since they have great social relevance. It is necessary to perform periodic longitudinal studies about the test use and its implications, since social and educational conditions in the context of the applicant population are dynamic.

**Keywords:** Higher education admission; Summative assessment; Multiple choice test; Student selection; Validity.

---

\*Contacto: melchorsm@unam.mx

issn: 1989-0397

www.rinace.net/riee/

https://revistas.uam.es/riee

Recibido: 4 de mayo de 2020

1ª Evaluación: 30 de junio de 2020

2ª Evaluación: 13 de julio de 2020

Aceptado: 21 de julio de 2020

## 1. Introducción

El presente trabajo ofrece una descripción del examen de admisión a las licenciaturas de la Universidad Nacional Autónoma de México (UNAM), la universidad más grande del país (Ordorika, Rodríguez y Montes de Oca, 2013) y la que recibe mayor cantidad de solicitudes de ingreso a nivel nacional (ANUIES, 2019). En concordancia con la importancia del proceso de selección para la institución y para la educación superior nacional, es menester someter a escrutinio los diversos elementos que constituyen el componente central del proceso, el examen escrito de conocimientos. A través del lente más importante en evaluación educativa, el de la validez (AERA, 2014; Kane, 2016; Shepard, 2016), en este trabajo se describen diversos atributos del instrumento, su diseño y los resultados de su aplicación, para identificar fortalezas y áreas de oportunidad en el instrumento mismo y su rol en el proceso de ingreso.

## 2. Fundamentación teórica

El marco teórico que sustenta este trabajo guarda relación con la naturaleza de los procesos de admisión a la educación superior, así como con la validez de las evaluaciones sumativas realizadas para este fin.

### *2.1. Procesos de admisión en educación superior a nivel internacional y nacional*

Las universidades que cuentan con procesos de selección, mediante los cuales analizan diversos elementos de los aspirantes para elegir periódicamente a las nuevas cohortes que ingresarán a sus espacios educativos, se enfrentan a un reto complejo y difícil. Los mecanismos específicos que utilizan las instituciones de educación superior para operacionalizar el proceso de selección varían entre países y universidades, dependiendo de varios factores: normatividad local, regional y nacional, carácter público o privado de la universidad, tamaño de la institución, población a quienes está dirigida, entre otros atributos (Manzi et al., 2010; Patterson et al., 2018; Trost, 1993).

Estos procesos frecuentemente incluyen un examen sumativo de alto impacto, dirigido principalmente a evaluar el conocimiento sobre las áreas relevantes a la carrera que se pretende ingresar. En ocasiones se suplementa con otros elementos como entrevistas, pruebas psicológicas, antecedentes académicos, actividades extracurriculares, exámenes por instancias externas a la universidad, entre otros (Patterson et al, 2018; Trost, 1993; Zwick, 2006). Si bien la selección para ingresar a los planteles que ofertan educación superior tiene múltiples aristas sociales, éticas, económicas, humanas, políticas y afectivas, la mayoría de las instituciones educativas privilegian aspectos primordialmente académicos para elegir a sus estudiantes, por razones de índole práctico, tradición, factibilidad, y la evidencia publicada que establece una correlación importante entre el desempeño académico antes de ingresar con el desempeño durante la licenciatura (Frey y Detterman, 2003; Juarros, 2006; Manzi et al., 2010; Patterson et al., 2018; Sigal y Dávila, 2004).

En un mundo ideal, el proceso de admisión a la educación superior incorporaría todos los elementos disponibles, realizaría procesos libres de sesgos, evaluaciones que exploraran las dimensiones más importantes de cada uno de los aspirantes, con jueces imparciales que balancearan la información obtenida de manera integral, equitativa y ponderada, para así seleccionar a los “mejores” candidatos. Desafortunadamente este proceso ideal no ocurre en el mundo real. No hay cabida en las instituciones de educación superior para toda la

población, por lo que las universidades y los gobiernos se ven obligados a diseñar complejos esquemas de selección que sean aceptados por la sociedad, la institución y los aspirantes, lo que hace inevitable que muchos aspirantes queden fuera de su primera elección (OCDE, 2018).

En algunos países existen exámenes nacionales estandarizados que se utilizan en el proceso de selección de las universidades, como el ACT y el SAT en los Estados Unidos, que proveen un elemento común de decisión a los organismos responsables del ingreso a las universidades (Frey y Detterman, 2003). En el caso de los Estados Unidos, el *Educational Testing Service* es una organización grande sin fines de lucro que se encarga de desarrollar exámenes estandarizados, con expertos que generan investigación original sobre el desarrollo, aplicación e interpretación de exámenes (Bennett, 2005). Estas organizaciones publican investigación original sobre sus instrumentos de medición en la literatura internacional, lo que tiene varios efectos: contribuye al conocimiento en evaluación educativa, legitima el uso de los instrumentos ante la comunidad académica y la sociedad, y genera una plataforma de evidencia de validez y confiabilidad de sus exámenes a partir de la cual pueden mejorarlos.

En el caso de nuestro país contamos con organizaciones similares (como el CENEVAL), que, si bien tienen expertos en evaluación educativa, se dedican principalmente a proveer un servicio y en menor medida a publicar en la literatura académica con arbitraje por pares trabajos de investigación sobre el tema, específicamente de la evidencia de validez de sus instrumentos (Gago, 2000). Las organizaciones de este tipo en Latinoamérica publican una abundante cantidad de manuales, reportes e informes externos, aunque las publicaciones sobre las evidencias de validez de sus instrumentos son pocas y se infiere que la validez de los mismos está documentada a través de mecanismos internos de control de calidad y los reportes técnicos correspondientes.

## ***2.2. Naturaleza sumativa de los procesos de admisión y sus instrumentos***

Los exámenes de admisión a las universidades se consideran evaluaciones sumativas de alto impacto o de altas consecuencias, ya que tienen potencial de generar efectos importantes en las personas que los toman (Cizek, 2001; Lane et al., 2016; Sánchez-Mendiola y Delgado-Maldonado, 2017). Estos efectos son económicos, sociales, educativos, e incluso en la salud física y mental de los sustentantes y sus familiares. También generan consecuencias inesperadas, positivas y negativas, lo que nos obliga a analizarlos con rigor. Por su naturaleza sumativa, la información obtenida de su aplicación y el análisis de sus resultados se mantienen en secreto como información reservada, pocas veces se divulgan en la literatura académica, y los reportes que reciben los sustentantes son escuetos y en ocasiones solo se les informa si acreditan o no el examen. El resultado es una ausencia de publicaciones que muestren con claridad y sustento metodológico el rigor académico de la elaboración de los instrumentos, así como del análisis de sus resultados, por lo que persiste la controversia sobre su utilidad real y sus implicaciones educativas (Martínez-Rizo, 2001; Sánchez-Mendiola y Delgado-Maldonado, 2017; Zwick, 2006).

## ***2.3. La Universidad Nacional Autónoma de México: la encrucijada del proceso de ingreso***

Al revisar la literatura latinoamericana sobre exámenes de admisión a la universidad, encontramos pocos estudios sobre las características psicométricas y evidencias de validez del uso de instrumentos de admisión (Backhoff, Tirado y Larrazolo, 2001; Buendía y

Rivera, 2010). Hasta la fecha no se ha publicado un análisis del examen de ingreso a las licenciaturas de la UNAM, a pesar de que es uno de los exámenes sumativos de alto impacto más importantes del país (Sánchez-Mendiola, 2017). Consideramos que es pertinente que la comunidad académica conozca los atributos principales del instrumento, los fundamentos conceptuales y metodológicos que sustentan su elaboración, análisis y control de calidad, para identificar áreas de oportunidad de mejora. De los aspirantes que presentan el examen de admisión a la Universidad Nacional Autónoma de México (UNAM), ingresan menos del 10%, lo que lo convierte en uno de los exámenes más selectivos del país (Guzmán y Serrano, 2011).

#### ***2.4. Evolución del concepto de validez y su valor como lente de análisis***

Para que los resultados de los procesos de evaluación tengan un robusto sustento y se utilicen de forma apropiada, es indispensable abordarlos desde la lente de la validez. Validez de un proceso de evaluación es el grado con el que mide lo que se supone que mide, tradicionalmente se clasificaba como las tres C: de contenido, de criterio y de constructo (Buntis, Buntis y Eggert, 2017; Sánchez-Mendiola, 2015; Young et al., 2018). En las últimas décadas el concepto ha evolucionado a una definición más amplia (AERA, 2014; Gregory, 2016; Kane, 2016; Kane y Bridgeman, 2017; Shepard, 2016). Actualmente se considera que se trata de un juicio valorativo holístico, en el que toda la validez es validez de constructo que se alimenta de diferentes fuentes. El concepto intenta responder a la pregunta: ¿qué inferencias pueden hacerse sobre la persona basándose en los resultados del examen? (Kane, 2016; Mendoza, 2015). No es el instrumento el que es válido *per se*, ya que la validez de un examen es específica para un propósito y se refiere más bien a lo apropiado de la interpretación de los resultados. En otras palabras, la validez no es una propiedad intrínseca del examen, sino del significado de los resultados en el entorno educativo específico y las inferencias que pueden hacerse de los mismos, por lo que el término “el instrumento es válido” es incorrecto (Kane, 2016; Sánchez, 2015).

Este modelo se ha construido a partir de las aportaciones metodológicas y conceptuales de varios autores como Messick y Kane, y si bien no está exento de controversia, es el actualmente aceptado por las principales organizaciones de evaluación educativa del mundo (AERA, 2014; Kane, 2016; Shepard, 2016).

### **3. Objetivo del estudio**

Se analizó el proceso de elaboración, análisis y control de calidad del examen de ingreso a las licenciaturas de la UNAM en su versión de febrero 2019, para mostrar aspectos técnico-metodológicos que puedan ser de utilidad para el desarrollo y análisis de este tipo de evaluaciones. Las expectativas del estudio fueron que, al seguir el proceso de elaboración del examen, pudieran obtenerse evidencias de validez que ofrecieran un panorama amplio de sus resultados.

### **4. Material y método**

Los diferentes elementos metodológicos que se utilizaron en el estudio se describen a continuación.

#### **4.1 Contexto**

Los procesos de selección de aspirantes a la educación superior tienen diversos componentes y fases, que reflejan las prioridades y realidades de cada institución educativa en su contexto (OCDE, 2018). En el caso de la UNAM se trata de un examen escrito de conocimientos. Desde 1997, la Dirección General de Evaluación Educativa de la UNAM formalizó la planeación general, definición del contenido y especificaciones de la prueba de admisión, atendiendo al Reglamento General de Inscripciones: “*Para ingresar a la Universidad es indispensable ser aceptado mediante concurso de selección, que comprenderá una prueba escrita y que deberá realizarse dentro de los periodos que al efecto se señalen*” (UNAM, 1997). La dependencia actualmente a cargo de la elaboración del examen es la Coordinación de Desarrollo Educativo e Innovación Curricular (CODEIC) de la UNAM, a través de la Dirección de Evaluación Educativa (Graue, 2018). La Dirección General de Administración Escolar expide las convocatorias para Concurso de Selección en febrero y junio de cada año, para aspirantes a ingresar al nivel Licenciatura en el Sistema Escolarizado y en el Sistema Universidad Abierta y Educación a Distancia (SUAYED) -modalidades Abierta y a Distancia-. Esta dependencia se encarga de la administración y logística del examen en las diversas sedes en que se aplica (DGAE, 2020).

#### **4.2 Diseño de investigación y marco conceptual**

Utilizamos el modelo de “la brújula de la investigación en educación” de Ringsted, Hodges y Scherpbier (2011), basado en los estudios de clasificación de investigación educativa de Cook, Bordage y Schmidt (2008). El centro de este modelo es un marco conceptual teórico, que en este estudio es el modelo de validez de Messick y Kane (Kane, 2016), adoptado por la *American Educational Research Association*, *American Psychological Association* y el *National Council of Measurement in Education* (AERA, 2014). En el modelo de la “brújula” de Ringsted hay cuatro categorías de estudios en educación, nuestro estudio encaja en la categoría de estudios exploratorios, enfocados en identificar y explicar fenómenos y sus relaciones, dentro del subtipo de estudios psicométricos que pretenden establecer evidencia de validez y confiabilidad de instrumentos de medición educativa (Ringsted, Hodges y Scherpbier, 2011).

#### **4.3 Metodología de elaboración del instrumento**

El marco conceptual de exámenes objetivos que utilizamos es el modelo del proceso de desarrollo y validación de exámenes de Haladyna y Downing (Lane et al., 2016). Este marco de desarrollo de exámenes objetivos es uno de los más utilizados en el mundo, se integra de 12 componentes y se apoya en los Estándares para Pruebas Educativas y Psicológicas de la AERA-APA-NCME (Lane et al., 2016):

- Componente 1. Plan general y global del examen
- Componente 2. Definición del dominio y declaraciones que se harán sobre los resultados
- Componente 3. Especificaciones del examen
- Componente 4. Desarrollo de los ítems
- Componente 5. Diseño y montaje del examen
- Componente 6. Producción del examen
- Componente 7. Aplicación del examen
- Componente 8. Calificación del examen

- Componente 9. Establecimiento de punto de pase
- Componente 10. Reporte de resultados del examen
- Componente 11. Seguridad del examen y banco de reactivos
- Componente 12. Reporte técnico de la prueba

En el caso del examen de la UNAM, el Componente 9 (establecimiento de punto de pase) no se realiza ya que la interpretación del examen es de índole normativa, no criterial, está sujeto principalmente al límite de espacios en la universidad en virtud de la excesiva demanda de aspirantes.

El examen es un instrumento de evaluación del conocimiento compuesto de reactivos de selección de respuesta con cuatro opciones, una de ellas correcta. La secuencia de desarrollo del instrumento se muestra en la figura 1 (AERA, 2014; Haladyna, Downing y Rodriguez, 2002; Lane et al., 2016).

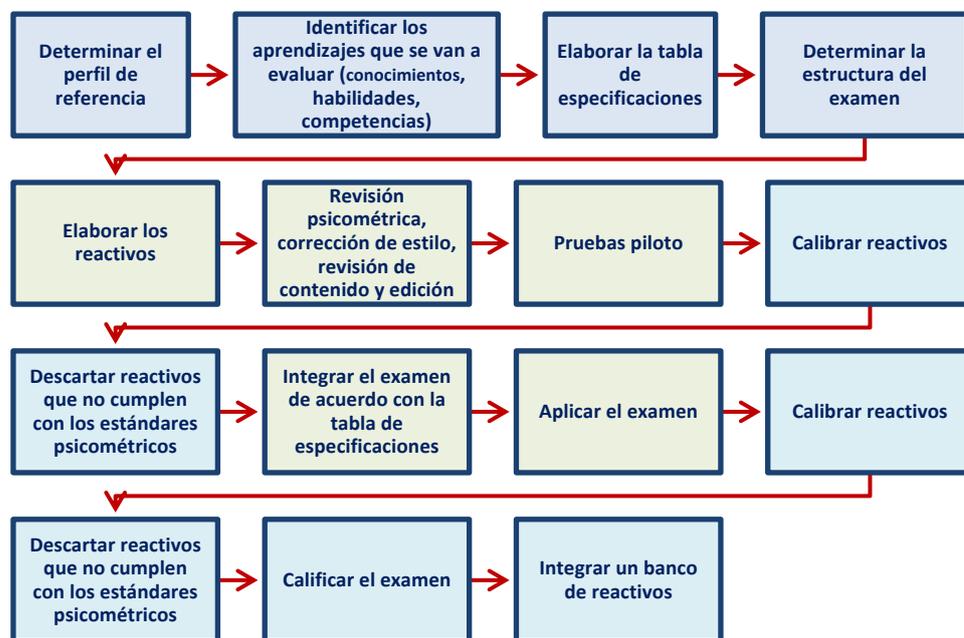


Figura 1. Metodología del diseño de exámenes de ingreso de la Universidad Nacional Autónoma de México

Fuente: Dirección de Evaluación Educativa de la CODEIC. UNAM.

En el modelo actual de validez, existen cinco fuentes importantes de la misma: contenido, procesos de respuesta, estructura interna (que incluye la confiabilidad y el comportamiento estadístico de los reactivos), relación con otras variables y consecuencias (AERA, 2014).

#### 4.4 Análisis psicométrico de reactivos

Se realizó el análisis psicométrico de reactivos con los modelos de la teoría de medición clásica (TMC) y de la teoría de respuesta al ítem (TRI) de uno, dos y tres parámetros (Andrich y Marais, 2019; Raykov y Marcoulides, 2016). El análisis con la TMC se realizó con el programa IteMan versiones 3.5 y 4 para el modelo de un parámetro, y BILOG-MG 3 para los modelos de dos y tres parámetros. El análisis de Rasch (Andrich y Marais, 2019; Boone y Noltemeyer, 2017) con el modelo de un parámetro específica que la

probabilidad de que un examinado  $i$  con habilidad  $\theta_i$  genere una respuesta  $x_{ij}$  al reactivo  $j$  con una dificultad  $b_j$  de acuerdo con la siguiente fórmula:

$$P(x_{ij}|\theta_i, b_j) = \frac{\exp(\theta_i - b_j)}{1 + \exp(\theta_i - b_j)}$$

Donde:

$x_{ij}$  = Respuesta del  $i$  – ésimo examinado al  $j$   
 – ésimo reactivo (1 si es correcto; 0 en otro caso).

$\theta_i$  = Habilidad del  $i$  – ésimo examinado.

$b_j$  = Dificultad del  $j$  – ésimo reactivo.

En el modelo de TRI de dos parámetros se usó la siguiente fórmula:

$$P(x_{ij}|\theta_i, b_j) = \frac{\exp [Da_j(\theta_i - b_j)]}{1 + \exp [Da_j(\theta_i - b_j)]}$$

Donde:

$x_{ij}$  = Respuesta del  $i$  – ésimo examinado al  $j$   
 – ésimo reactivo (1 si es correcto; 0 en otro caso).

$\theta_i$  = Habilidad del  $i$  – ésimo examinado.

$a_j$  = Discriminación del  $j$  – ésimo reactivo.

$b_j$  = Dificultad del  $j$  – ésimo reactivo.

$D$  = Factor de escala para aproximar la función a una ojiva normal (1.7).

Para el modelo de TRI de tres parámetros se utilizó la siguiente fórmula:

$$P(x_{ij}|\theta_i, b_j) = c_j + (1 - c_j) \frac{\exp [Da_j(\theta_i - b_j)]}{1 + \exp [Da_j(\theta_i - b_j)]}$$

Donde:

$x_{ij}$  = Respuesta del  $i$  – ésimo examinado al  $j$   
 – ésimo reactivo (1 si es correcto; 0 en otro caso).

$\theta_i$  = Habilidad del  $i$  – ésimo examinado.

$a_j$  = Discriminación del  $j$  – ésimo reactivo.

$b_j$  = Dificultad del  $j$  – ésimo reactivo.

$c_j$  = pseudo oportunidad (adivinación) del  $j$  – ésimo reactivo.

$D$  = Factor de escala para aproximar la función a una ojiva normal (1.7).

Se realizó análisis del funcionamiento diferencial de los ítems por sexo (DIF, por sus iniciales en inglés), de acuerdo con el procedimiento de Mantel-Haenszel con la modificación del modelo de Rasch de un parámetro (García-Medina, Martínez-Rizo y Cordero Arroyo, 2016; Linacre y Wright, 1989). Se utilizaron los criterios del ETS para identificar DIF insignificante, leve-moderado y moderado-alto (Dorans y Holland, 1992)

## 5. Resultados

El examen se aplicó con la metodología descrita para exámenes sumativos de gran escala (Lane et al., 2016), en instrumentos impresos a contestar con lápiz y hoja de respuestas, los días 23 y 24 de febrero de 2019, en 25 sedes en el área metropolitana de la Ciudad de México. El examen fue respondido por 148.407 aspirantes a las licenciaturas de las cuatro áreas de conocimiento que oferta la UNAM: CFMI=Ciencias Físico Matemáticas e Ingenierías; CBQS=Ciencias Biológicas, Químicas y de la Salud; CS=Ciencias Sociales; HyA=Humanidades y Artes (cuadro 1).

Cuadro 1. Número y porcentaje de aspirantes por área del conocimiento

ÁREA	N	%
CFMI	30.561	20,6
CBQS	56.474	38,1
CS	45.229	30,5
HyA	16.143	10,9
Total	148.407	100,0

CFMI=Ciencias Físico Matemáticas e Ingenierías; CBQS=Ciencias Biológicas, Químicas y de la Salud; CS=Ciencias Sociales; HyA=Humanidades y Artes.

Fuente: Elaboración propia con información de la Dirección General de Administración Escolar (DGAE) de la UNAM.

En el cuadro 2 se muestra el número y porcentaje de aspirantes por sexo y área del conocimiento solicitada.

Cuadro 2. Número y porcentaje de aspirantes por sexo y área de conocimiento

	ÁREA				TOTAL
	CFMI	CBQS	CS	HYA	
Hombres	21.486	18.082	20.723	5.638	65.929
%	32,6%	27,4%	31,4%	8,6%	100,0%
% en el área	70,3%	32,0%	45,8%	34,9%	44,4%
Mujeres	9.075	38.392	24.506	10.505	82.478
%	11,0%	46,5%	29,7%	12,7%	100,0%
% en el área	29,7%	68,0%	54,2%	65,1%	55,6%
Total	30.561	56.474	45.229	16.143	148.407
%	20,6%	38,1%	30,5%	10,9%	100,0%

CFMI=Ciencias Físico Matemáticas e Ingenierías; CBQS=Ciencias Biológicas, Químicas y de la Salud; CS=Ciencias Sociales; HyA=Humanidades y Artes.

Fuente: Elaboración propia con información de la Dirección General de Administración Escolar (DGAE) de la UNAM.

Se analizó la información del examen que atendiera a cada una de las fuentes de evidencia de validez descritas en la sección de Método, que a continuación se describen:

- Contenido.* El contenido del examen se fundamentó en los planes de estudio de la educación media superior. Se estableció un perfil de referencia por cuerpos colegiados universitarios, posteriormente comisiones de profesores del bachillerato de la UNAM, expertos en contenido, revisaron los temarios y determinaron los temas y niveles cognitivos a evaluar. Se elaboró una tabla de especificaciones con los resultados de aprendizaje esperados y se ponderaron las áreas del conocimiento a explorar. Los académicos elaboradores de reactivos fueron entrenados para elaborar preguntas de opción múltiple de características técnicas apropiadas. La estructura del examen se muestra en el cuadro 3, integrándose con 120 reactivos.

Cuadro 3. Número de reactivos del examen de ingreso a las licenciaturas de la UNAM, por área del conocimiento

MATERIA	CFMI	CBQS	CS	HyA
Matemáticas	26	24	24	22
Física	16	12	10	10
Química	10	13	10	10
Biología	10	13	10	10
Historia universal	10	10	14	10
Historia de México	10	10	14	10
Literatura	10	10	10	10
Geografía	10	10	10	10
Español	18	18	18	18
Filosofía	-	-	-	10
Total	120	120	120	120

CFMI=Ciencias Físico Matemáticas e Ingenierías; CBQS=Ciencias Biológicas, Químicas y de la Salud; CS=Ciencias Sociales; HyA=Humanidades y Artes.

Fuente: Elaboración propia con información de la Dirección de Evaluación Educativa de la CODEIC UNAM.

- Procesos de respuesta.* Este apartado se refiere a evidencia de integridad de los datos de manera que las fuentes de error que se pueden asociar con la administración del examen han sido controladas en la medida de lo posible. Uno de ellos es la familiaridad del estudiante con el formato de preguntas de opción múltiple, lo cual se cumple en la actualidad. Al ser con lápiz y papel no introduce la variable de habilidad en el uso de computadoras. Cada reactivo es revisado por personal técnico para verificar congruencia, relación con el resultado de aprendizaje y estructura gramatical. Se efectúa la validación de la clave de respuestas, así como el control de calidad del reporte de resultados.

Desarrollamos una plataforma informática para el desarrollo y validación del examen, que tiene más de una década de perfeccionamiento e integración, el “Sistema Integral de Gestión de Exámenes” (SIGE UNAM, marca registrada), lo que proporciona un elemento más de validez al desarrollo del examen, la validación de los reactivos, y la integración del banco de reactivos con características apropiadas. En el SIGE se capturan y validan los reactivos por tres expertos en contenido, y transitan por el proceso de corrección de estilo, inclusión de los resultados de aprendizaje, entre otros aspectos técnicos. En el sistema se captura el historial del desempeño psicométrico del reactivo.

- Estructura interna.* Se refiere a las características estadísticas del examen, como estadísticas descriptivas y análisis de reactivos, el funcionamiento de los distractores, la confiabilidad del examen, entre otros (AERA, 2014). En la figura 2

se muestra la distribución de aciertos por área del conocimiento, en la que se observa claramente una tendencia de agrupamiento de los aspirantes hacia la izquierda en el área de menor cantidad de aciertos. Los patrones de distribución de aciertos en las versiones y ordenamientos del examen son prácticamente idénticas (datos no mostrados).

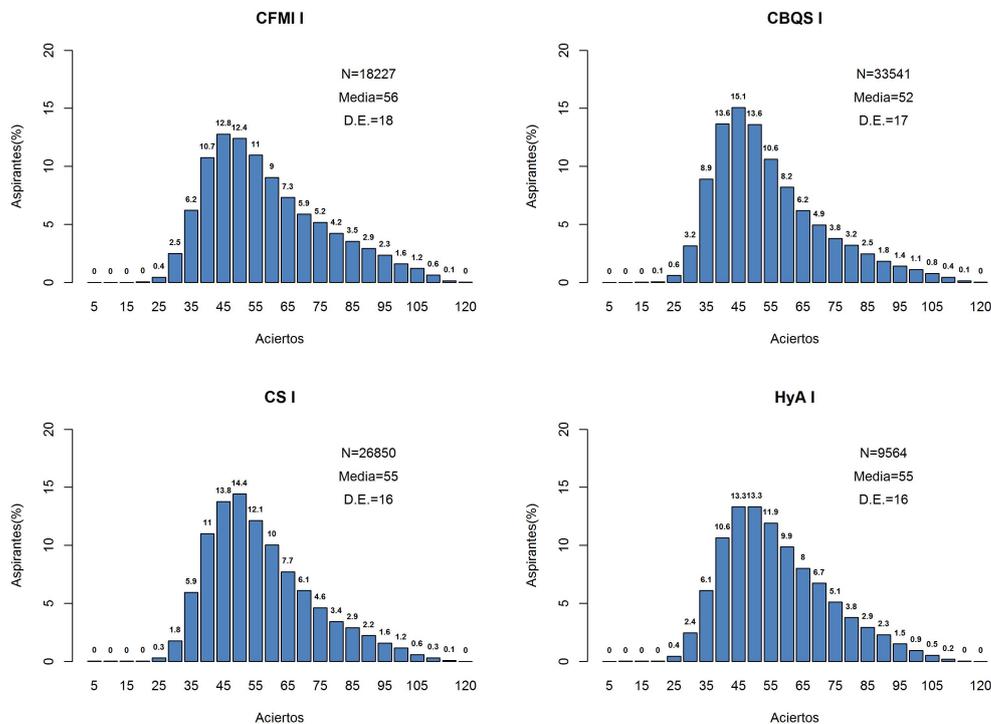


Figura 2. Distribución de aciertos en el examen de admisión a la licenciatura de la UNAM, por área del conocimiento  
 CFMI=Ciencias Físico Matemáticas e Ingenierías; CBQS=Ciencias Biológicas, Químicas y de la Salud; CS=Ciencias Sociales; HyA=Humanidades y Artes.

Fuente: Elaboración propia.

En el cuadro 4 se presentan los resultados del análisis psicométrico con la TMC, incluyendo índice de dificultad promedio, índice de discriminación promedio con coeficiente de punto biserial, error estándar de medición y confiabilidad determinada con el alfa de Cronbach.

Cuadro 4. Análisis psicométrico con la Teoría de Medición Clásica del examen de ingreso a la licenciatura de la UNAM, por campo de conocimiento y versión del examen

	CFMI		CBQS		CS		HYA	
	I	II	I	II	I	II	I	II
N	18.227	12.334	33.541	22.933	26.850	18.379	9.564	6.579
Promedio de aciertos	55,4	55,7	52,5	52,9	54,8	54,6	54,8	55,1
Desviación estándar	18,2	18,1	17,0	16,9	16,3	16,4	16,3	16,6
Mediana	52	52	49	49	52	51	52	52
EEM	4,92	4,92	4,98	5,00	4,97	4,98	4,99	4,97
Dificultad media	0,466	0,468	0,437	0,441	0,456	0,455	0,456	0,460
CPB media	0,298	0,296	0,273	0,271	0,260	0,263	0,258	0,264
Alfa de Cronbach	0,927	0,926	0,914	0,913	0,907	0,908	0,906	0,910

N=148,407. CFMI=Ciencias Físico Matemáticas e Ingenierías; CBQS=Ciencias Biológicas, Químicas y de la Salud; CS=Ciencias Sociales; HyA=Humanidades y Artes; EEM=error estándar

de medición; CPB=correlación punto biserial.

Fuente: Elaboración propia.

En el cuadro 5 se presentan los resultados del análisis psicométrico con la TRI.

Cuadro 5. Análisis psicométrico con la Teoría de Respuesta al Ítem de uno, dos y tres parámetros, del examen de ingreso a la licenciatura de la UNAM, por campo de conocimiento

Área	r	Modelos de TRI											
		Un parámetro			Dos parámetros				Tres parámetros				
		b			a	b		a	b	c			
$\bar{x}$	D.E.	$\bar{x}$	D.E.	$\bar{x}$	D.E.	$\bar{x}$	D.E.	$\bar{x}$	D.E.	$\bar{x}$	D.E.	$\bar{x}$	D.E.
CFMI	204	0,277	1,053	0,788	0,381	0,540	1,968	1,409	0,842	1,009	1,745	0,189	0,089
CBQS	201	0,455	1,012	0,710	0,313	0,616	1,429	1,458	0,777	1,073	0,960	0,200	0,116
CS	204	0,253	1,155	0,703	0,336	0,303	1,963	1,305	0,776	0,830	1,229	0,197	0,116
HyA	200	0,311	1,195	0,715	0,346	0,790	3,947	1,136	0,604	1,021	1,799	0,181	0,094
Total	809	0,324	1,109	0,729	0,347	0,561	2,517	1,327	0,765	0,983	1,479	0,192	0,105

N=148,407. r=reactivos únicos por área; TRI=Teoría de Respuesta al Ítem; D.E.=Desviación estándar; CFMI=Ciencias Físico Matemáticas e Ingenierías; CBQS=Ciencias Biológicas Químicas y de la Salud; CS=Ciencias Sociales; HyA=Humanidades y Artes.

Fuente: Elaboración propia.

Todos los reactivos que se utilizan en el examen cubren criterios psicométricos predefinidos, obtenidos con el análisis psicométrico de TMC y TRI. Todos provienen de un banco de reactivos extenso, y han sido analizados en estudios piloto con estudiantes para documentar su comportamiento psicométrico previo. En la figura 3 podemos ver los resultados del mapa de Wright, comparando la dificultad de los reactivos con la habilidad de los aspirantes, en el área de las Ciencias Físico Matemáticas e Ingenierías, con el modelo de Rasch de la TRI (Andrich y Marais, 2019; Boone y Noltemeyer, 2017). Los resultados en las otras tres áreas del conocimiento presentaron patrones similares, lo que arroja evidencia de validez sobre lo apropiado de la dificultad del examen para el rango de niveles de habilidad de los aspirantes. Es importante notar que todos los sustentantes están incluidos en el rango de dificultad del instrumento, y que de manera similar a la distribución estadística del número de aciertos en la figura 2, hay mayor concentración de estudiantes hacia el extremo de menor habilidad.

A partir de 2018, comenzamos a realizar análisis diferencial de los ítems (DIF, por sus siglas en inglés), para explorar el comportamiento del examen y los reactivos por sexo, tema que ha sido sujeto de constante debate (Guzmán y Serrano, 2011). Un reactivo presenta DIF cuando los examinados de *un mismo nivel de habilidad*, pero provenientes de diferentes grupos, tienen una probabilidad distinta de contestarlo correctamente (Walker, 2011; Zieky, 1993). Se empleó la técnica de TRI basada en el modelo de Rasch, donde los grupos de interés fueron las mujeres y los hombres. Si el contraste en el nivel de dificultad de un reactivo entre los grupos de interés no supera los 0,43 lógitos representa un DIF sin importancia; si es mayor a 0,64 lógitos representa DIF moderado-alto; si se encuentra entre estos valores se trata de un DIF leve-moderado (Holland y Weiner 1993). Encontramos muy pocos reactivos con DIF leve-moderado, los cuales fueron valorados por un cuerpo colegiado para analizar la lógica del reactivo y el resultado de aprendizaje explorado, y así determinar su potencial efecto en los resultados. En la figura 4 podemos ver un ejemplo de los reactivos del área de Matemáticas, en los que no se encontraron reactivos con DIF en cuanto a sexo.

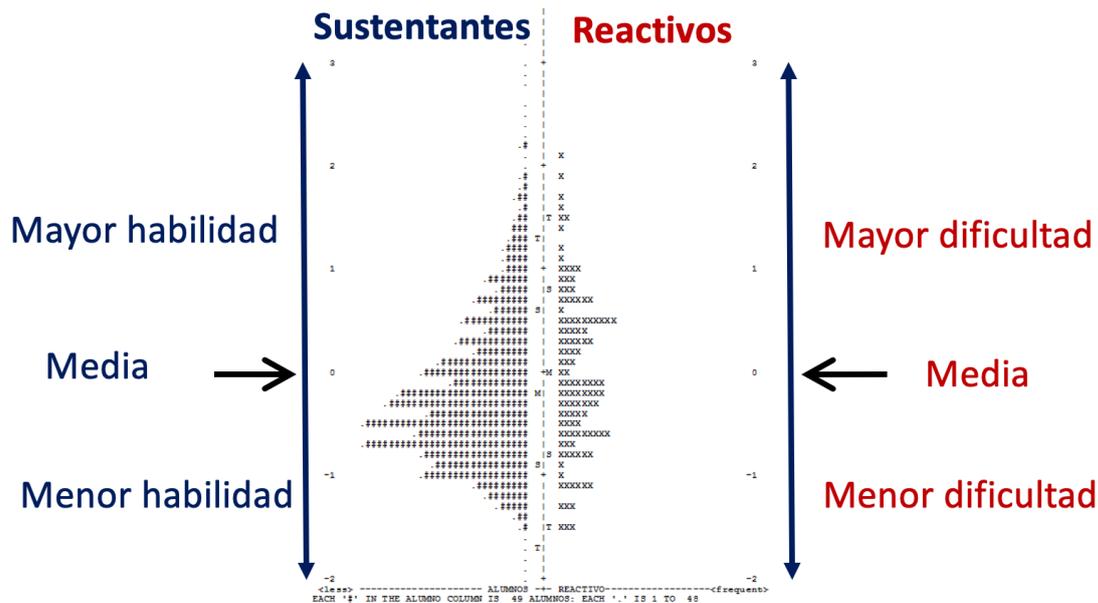


Figura 3. Mapa de dificultad de los reactivos y habilidad de los estudiantes, con el Modelo de Rasch  
 N=148.407  
 Fuente: Elaboración propia.

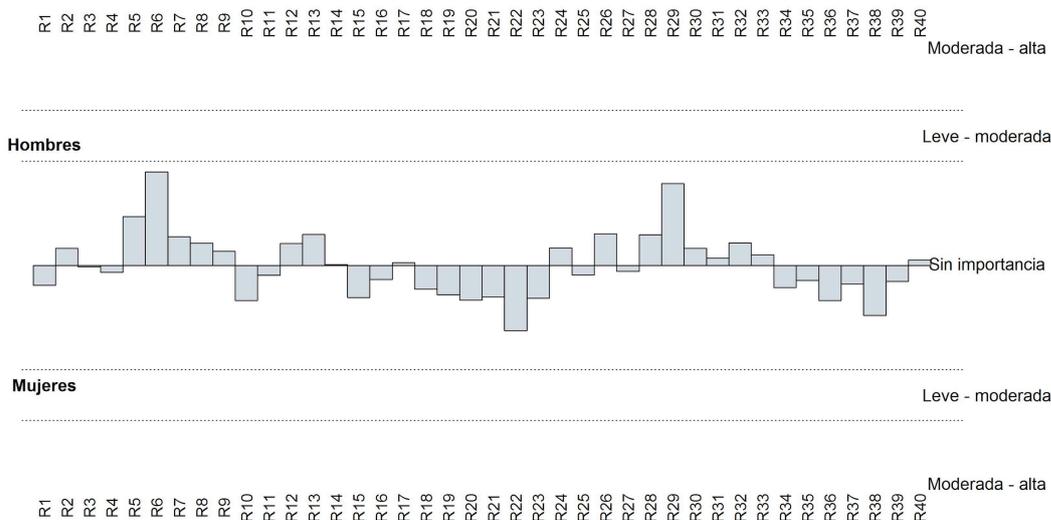


Figura 4. Funcionamiento diferencial de los ítems (DIF) de los reactivos de Matemáticas del examen de admisión a la licenciatura de la UNAM de febrero de 2019 según el sexo de los aspirantes  
 Fuente: Elaboración propia.

En la figura 5 podemos observar un ejemplo de la curva de un reactivo con el modelo de Rasch, con DIF mínimo o intrascendente.

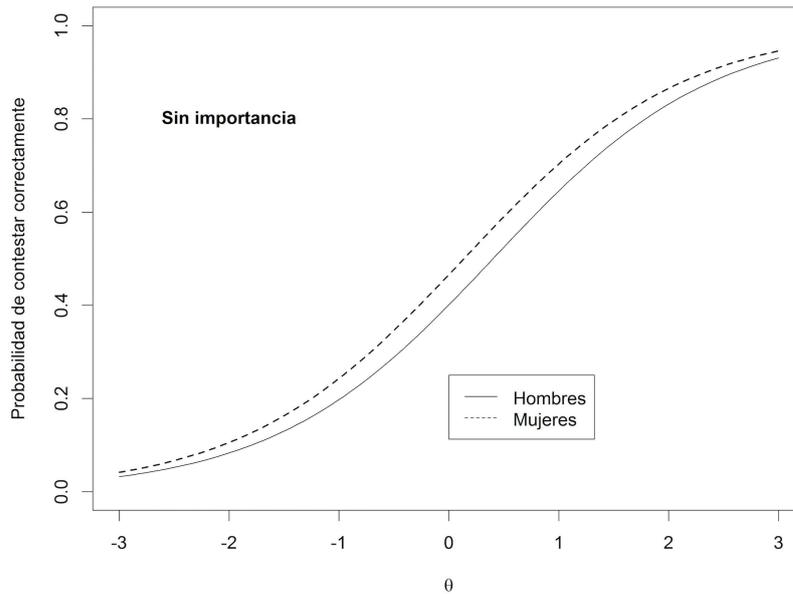


Figura 5. Visualización de un reactivo de matemáticas del examen de admisión a la licenciatura de la UNAM con un funcionamiento diferencial (DIF) sin importancia por sexo  
Fuente: Elaboración propia.

En la figura 6 se observa el DIF por sexo de los reactivos de Español, del área de Ciencias Biológicas, Químicas y de la Salud. Muy pocos tienen DIF leve-moderado, por lo que fueron evaluados para determinar el potencial impacto en el examen. Como ocurre en exámenes de este tipo, la dirección de algunos reactivos con DIF leve para hombres se cancela con los reactivos con DIF leve para mujeres. Es importante destacar que estas cifras son solamente elementos estadísticos, que por sí solos no documentan sesgo a favor o en contra de una población, deben evaluarse cualitativamente por un grupo de expertos en contenido y evaluación, para analizar su potencial efecto en los resultados.

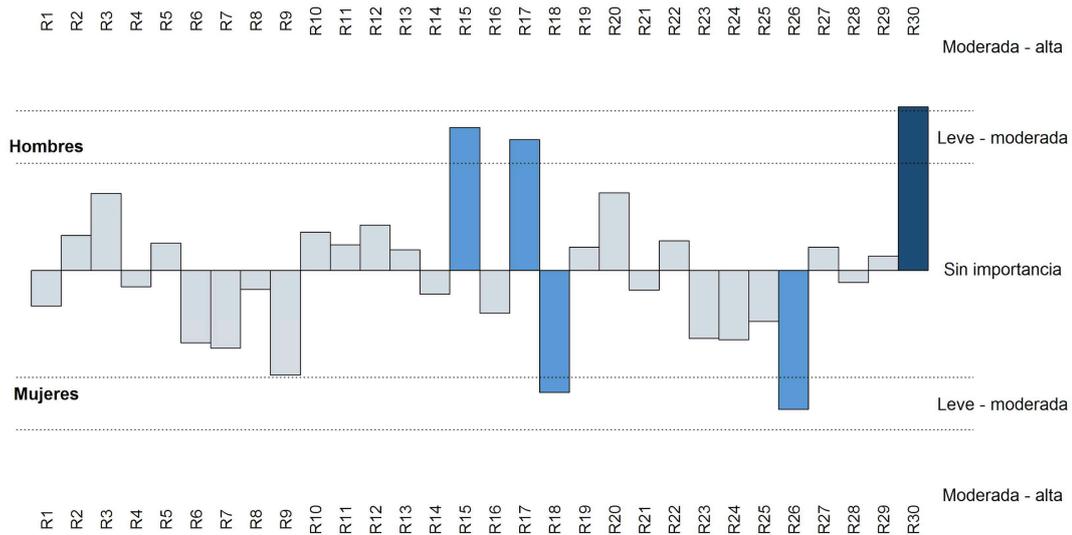


Figura 6. DIF por sexo de los reactivos de Español, del área de Ciencias Biológicas, Químicas y de la Salud, en el examen de admisión a la licenciatura de la UNAM, febrero de 2019  
Fuente: Elaboración propia.

- *Relación con otras variables:* La relación de los resultados del examen con otras variables se refiere a la correlación estadística entre los resultados obtenidos en el mismo con otra medición de características conocidas. El examen de ingreso a la UNAM se asocia con los resultados de los exámenes de diagnóstico de conocimientos que se aplica a todos los estudiantes que ingresan a la institución ( $r=0,64$ ,  $p<0,01$ ). Las correlaciones por área del conocimiento son similares. Actualmente estamos explorando la correlación del examen de admisión con el desempeño escolar a lo largo de las carreras y la eficiencia terminal. Por otra parte, encontramos una importante relación entre el desempeño en el examen diagnóstico al ingreso de la UNAM y el éxito en sus trayectorias académicas (Martínez-González et al., 2018). Al existir correlación entre el examen de ingreso y el examen diagnóstico, y entre el examen diagnóstico y el éxito académico, pudiera existir correlación entre el examen de ingreso y el desempeño a lo largo de la carrera, hipótesis que debe probarse.
- *Consecuencias:* Se refiere al impacto en los estudiantes de las puntuaciones de la evaluación, de las decisiones que se toman como resultado del examen, y su efecto en la enseñanza y el aprendizaje. Por ejemplo: el método de establecimiento del punto de corte, las consecuencias para el estudiante y la sociedad, las consecuencias para los profesores y las instituciones educativas. En este apartado no contamos con fuentes de evidencia de validez propias, por lo que hay un espacio de oportunidad amplio para realizar estudios de los costos económicos y emocionales, costos sociales de falsos positivos y falsos negativos, entre otros aspectos del proceso de selección.

## 6. Discusión y conclusiones

El análisis del examen de ingreso a las licenciaturas de la UNAM ofrece un panorama de información con datos de diversas fuentes de evidencia de validez, que proveen un retrato evaluativo del instrumento. Validez en evaluación educativa implica una aproximación científica a la interpretación de los resultados de los exámenes, es decir, probar hipótesis sobre los conceptos evaluados en el examen (AERA, 2014; Kane, 2016; Shepard, 2016). La información proporcionada por el uso de un instrumento de evaluación no es válida o inválida *per se*, sino que forma parte de un espectro de datos e información que deben utilizarse de manera sensata, integral y contextualizada, en el sentido de que las puntuaciones obtenidas por los sustentantes en un examen proveen más o menos evidencia para apoyar o rechazar una interpretación específica (por ejemplo aprobar o no un curso, admitir o no a un estudiante en la universidad) (Buntis, Buntis y Eggert, 2017; Kane, 2016; Mislevy, 2016; Young et al., 2016). Las organizaciones que elaboran e implementan los exámenes sumativos de alto impacto (entidades gubernamentales, instituciones educativas) son los principales responsables de validar las afirmaciones que hacen sobre la interpretación de los resultados de un examen, ya que generalmente son quienes tienen los elementos y recursos para hacerlo (AERA, 2014; Gregory, 2016; Sireci, 2016). Las dependencias universitarias que elaboran exámenes tienen la obligación ética de documentar qué tan defendible es la interpretación de los resultados.

Los resultados descritos en este trabajo proporcionan evidencias de validez para el uso de este examen como instrumento de evaluación del conocimiento. Los datos obtenidos y sus magnitudes son compatibles con los recomendados por organizaciones nacionales e

internacionales para exámenes sumativos de alto impacto a gran escala, desde la estrategia de elaboración del instrumento, hasta las cifras de análisis psicométrico y confiabilidad (AERA, 2014; Young et al., 2016). Es importante enfatizar que el examen de ingreso a la Universidad es solamente un instrumento de medición del conocimiento, nada más, pero tampoco nada menos. El uso de los resultados que se obtienen con los instrumentos de evaluación y las inferencias que se hacen de los mismos es un tema extraordinariamente complejo.

En las últimas décadas las principales organizaciones de evaluación educativa del mundo han hecho énfasis en la necesidad de que se incluyan elementos que propicien justicia y equidad en el proceso, para ser congruentes con el sentido social de la educación (AERA, 2014). Existe controversia sobre el tema, ya que los exámenes estandarizados en gran escala, que por necesidad utilizan instrumentos uniformes que se aplican en contextos altamente controlados, con la intención de que cada estudiante se enfrente al mismo reto en igualdad de condiciones, por definición tratan a todos los estudiantes de la misma manera. Por ejemplo, la riqueza de la heterogeneidad de los seres humanos es poco susceptible de “medirse” con instrumentos estandarizados, ya que estos no capturan fácilmente los matices de la individualidad de las personas. Además, los resultados individuales en un examen sumativo, en un momento específico en el tiempo, no necesariamente reflejan la realidad holística y longitudinal de la persona, ni de forma absoluta su nivel preciso de conocimiento y aplicación del mismo en solución de problemas en la vida real. Esta tensión entre las diversas perspectivas filosóficas de lo que debe ser la evaluación educativa continúa sin resolverse, lo que motiva discusiones intensas en contextos académicos (Martínez-Rizo, 2001; Sánchez-Mendiola y Delgado-Maldonado, 2017).

El hecho es que el ingreso de aspirantes a organizaciones que tienen recursos y cupo limitados, como las universidades públicas, obliga a tomar decisiones difíciles que no dejan totalmente satisfecha a la población, principalmente a aquellos que no son admitidos. La comprensión cabal del concepto moderno de validez por la comunidad académica es fundamental para entender las limitaciones de los resultados de los exámenes, ya que extrapolar conclusiones y decisiones más allá de lo académicamente sensato es inapropiado.

Es indispensable explorar diversos mecanismos tanto tradicionales como innovadores, que puedan combinarse para generar procesos de admisión socialmente aceptables, metodológicamente correctos y éticamente justificables, tarea compleja en la época actual, especialmente si los recursos son limitados. A nivel internacional el inexorable proceso de expansión de la demanda por la educación superior se ha asociado con respuestas diferenciadas en regiones, países y universidades (Sigal y Dávila, 2004; Trost, 1993). En algunos casos se realizan estrategias nacionales rigurosas con un mecanismo centralizado que usa diferentes elementos evaluativos como en Chile, en otros como Argentina se ha usado el acceso irrestricto, y en otros países como Colombia, Brasil y Panamá se emplean diferentes esquemas de exámenes de admisión, pruebas estandarizadas y una variedad de metodologías en sus procesos (Sigal y Dávila, 2004; Trost, 1993). En México cada institución educativa define sus mecanismos de ingreso, así como los criterios de selección y los niveles de exigencia en los diferentes componentes del proceso de acuerdo a su normatividad y criterios técnicos. Varias universidades mexicanas utilizan instrumentos estandarizados desarrollados por organizaciones dedicadas a evaluación educativa, como el Centro Nacional de Evaluación para la Educación Superior (CENEVAL, 2020), además

de de evaluaciones psicológicas, el promedio del bachillerato, entrevistas personales, ensayos, entre otros. Cada institución lo hace de acuerdo a sus recursos, normatividad, posibilidades y tamaño de la demanda. Es importante hacer notar que algunas instituciones que utilizan varios elementos e instrumentos de evaluación, generalmente no hacen público el proceso en detalle, como la ponderación de cada componente y cómo integran un resultado final que les permita tomar una decisión, por lo que el proceso se convierte en una especie de “caja negra” que puede dar lugar a desconfianza en el mismo.

El caso de la UNAM es único en el país, ya que es la institución de educación media superior y superior de México más reconocida a nivel nacional e internacional, y al ser pública y para fines prácticos gratuita, es solicitada por un número creciente de aspirantes. Actualmente la UNAM tiene 360.883 estudiantes (DGPL-UNAM, 2020) de los que 217.808 (60,3%) son de licenciatura. De acuerdo con la DGAE (2019), en el ciclo académico 2018-2019 participaron en el concurso de selección a las licenciaturas de la UNAM 261.157 aspirantes, de los cuales ingresaron 24.007 (9,2%). La UNAM tiene además la particularidad de que más de la mitad de sus estudiantes de ingreso a la licenciatura (aproximadamente el 55%) provienen de los bachilleratos de la misma institución, por el mecanismo de pase reglamentado. De cualquier manera, es necesario identificar las fuentes de evidencia de validez del examen de ingreso a las licenciaturas, ya de ello depende la decisión de ingreso por concurso de selección. El examen es el principal elemento decisorio, no se toma en cuenta el promedio en el bachillerato ni otros tipos de instrumentos externos, estudios psicológicos o entrevista personal, en virtud de la gran cantidad de aspirantes, la heterogeneidad de los promedios en las escuelas del sistema educativo nacional y las dificultades éticas y logísticas de utilizar otros elementos de forma equitativa y válida, como para incluirlos de manera ponderada en el puntaje de ingreso.

El examen tiene números satisfactorios desde el punto de vista psicométrico, incluyendo grado de dificultad, discriminación, confiabilidad, entre otros. Son pocos los estudios en nuestro país que documenten este tipo de cifras para establecer comparaciones (Backhoff, Tirado y Larrazolo, 2001), aunque en el diálogo con colegas de otras universidades refieren que sus exámenes tienen evidencia de validez, de acuerdo con sus reportes internos y mecanismos de control de calidad. En el contexto moderno de la educación superior en el que la demanda sobrepasa con mucho a la oferta, es importante que los exámenes estandarizados de alto impacto dejen de ser percibidos como un instrumento punitivo o de control, y que se haga conciencia de que es una herramienta académica que debe elaborarse con profesionalismo y atención a los detalles técnicos (AERA, 2014). Por otra parte, es necesario repensar el ingreso a la universidad como un sistema del que un examen escrito de conocimientos es un elemento del proceso, y reflexionar sobre el peso que se les da a los atributos exclusivamente académicos de los aspirantes (AERA, 2014; Juarros, 2006; Patterson et al., 2018).

El estudio tiene algunas limitaciones, ya que analiza solo una institución y los resultados pueden no ser aplicables en su totalidad a las demás universidades en México y Latinoamérica, por diferencias en tamaño, recursos y normatividad interna. A pesar de ello, creemos que al tratarse de la institución de educación superior más grande del país y una de las más grandes del mundo, con una gran cantidad de aspirantes y una baja tasa de aceptación, la descripción de la metodología y resultados encontrados es útil para la comunidad académica y tomadores de decisiones educativos, ya que presenta datos concretos y métodos rigurosos de medición educativa. Además, comienza a romper el paradigma tradicional de exceso de secrecía en los exámenes de gran escala y alto impacto,

para informar a la comunidad académica sobre estos temas. Generalmente la sociedad, los aspirantes y una parte del gremio docente desconocen los aspectos técnicos sofisticados de la medición educativa, mismos que dan fortaleza a los resultados del examen y el uso que se hace de ellos.

Es importante reconocer las limitaciones de los exámenes estandarizados de alto impacto, y la percepción social que se tiene de ellos en la actualidad. Este estudio se limitó a identificar las fuentes de evidencia de validez disponibles, pero no todas las que potencialmente existen, principalmente las relacionadas con consecuencias, costos sociales, psicológicos, económicos y de efectos a largo plazo en los aspirantes. Tampoco hemos analizado el seguimiento a largo plazo de los estudiantes, para identificar la correlación entre el examen de selección, la graduación y el éxito profesional de los graduados. Creemos que los espacios de oportunidad para ampliar nuestro conocimiento de los exámenes de alto impacto, y del proceso de selección en las universidades, es de gran importancia para las instituciones educativas y la sociedad en general.

Uno de los métodos más utilizados para tratar de identificar sesgos en los exámenes que pudieran poner en desventaja a algún grupo de personas, ya sea por sexo o nivel socioeconómico, es el análisis DIF (Alavi y Bordbar, 2017; Yavuz et al., 2018; Zieky, 1993). En nuestra institución iniciamos en 2018 este tipo de análisis por sexo y por bachillerato de procedencia (público o privado), y no identificamos niveles de DIF importante en los reactivos del examen que pudieran comprometer sus resultados e inferencias. El análisis detallado con esta metodología será sujeto de otra publicación, ya que incorporaremos en el análisis los niveles socioeconómicos, variable por demás importante en nuestro contexto. Existen muy pocos estudios publicados sobre el uso de esta metodología en el país (García-Medina et al., 2016), por lo que consideramos debe impulsarse su uso en las universidades que realicen exámenes sumativos de alto impacto.

Los mecanismos y políticas de admisión a la educación superior son el resultado de una convergencia de factores históricos, sociales y de disponibilidad de espacio y recursos a lo largo del tiempo, y cada país, región y universidad han enfrentado el reto de acuerdo con su realidad local. No parece que a corto plazo vaya a disminuir la demanda de espacios, y el crecimiento de las universidades existentes requiere ineludiblemente de más recursos utilizados de manera eficaz para dar respuesta a las necesidades sociales. Es inevitable que los aspirantes a ingresar a la educación superior deseen hacerlo a las universidades con mayor prestigio y que no impliquen gastos directos de bolsillo para transitar en la licenciatura, por lo que las solicitudes de la mayoría de los estudiantes se concentran en unas cuantas instituciones, principalmente públicas, para seguir adelante en su trayectoria de vida. Ello implica la necesidad mecanismos de selección que impactan a gran cantidad de aspirantes.

La responsabilidad de realizar buenos exámenes e informar a la sociedad sobre sus limitaciones recae en nuestras organizaciones y grupos de expertos, en colaboración con la comunidad académica, las autoridades y los medios de comunicación (AERA, 2014; Martínez-Rizo, 2016). La asimetría de poder intrínseca en los procesos de evaluación conlleva una enorme responsabilidad de las autoridades institucionales, por lo que estos procesos deben realizarse atendiendo el estado del arte de la evaluación educativa, y promoviendo la profesionalización en este tema de los participantes en el proceso de admisión. Es importante explorar mecanismos alternativos e identificar elementos de decisión que pudieran incorporarse al proceso de admisión en nuestras universidades, pero también es fundamental revisar rigurosamente los mecanismos de selección existentes y

mejorar los instrumentos a la luz de publicaciones como la presente. La investigación en evaluación educativa debe convertirse en una prioridad de nuestras instituciones de educación superior (Schuwirth et al., 2010). Las universidades, espacio de trabajo de académicos de alto nivel, deben proveer servicios de evaluación educativa acordes con su prestigio institucional.

## Referencias

- Alavi, S. y Bordbar, S. (2017). Differential item functioning analysis of high-stakes test in terms of gender: A Rasch model approach. *Malaysian Online Journal of Educational Sciences*, 5(1), 10-24.
- American Educational Research Association, American Psychological Association, National Council on Measurement in Education and Joint Committee on Standards for Educational and Psychological Testing. (2014). *Standards for educational and psychological testing*. AERA.
- Andrich, D. y Marais, I. (2019). A Course in Rasch Measurement Theory. En D. Andrich y I. Marais (Coords.), *Measuring in the Educational, Social and Health Sciences* (pp. 41-53). Springer.
- Asociación Nacional de Universidades e Instituciones de Educación Superior ANUIES. (2019). *Anuario estadístico de la población escolar en la educación superior. Técnico Superior y Licenciatura, ciclo 2017-2018*. Recuperado de <http://www.anuiemx.com/informacion-y-servicios/informacion-estadistica-de-educacion-superior/anuario-estadistico-de-educacion-superior>
- Backhoff, E., Tirado, F. y Larrazolo, N. (2001). Ponderación diferencial de reactivos para mejorar la validez de una prueba de ingreso a la universidad. *Revista Electrónica de Investigación Educativa*, 3(1), 1-10.
- Bennett, R. E. (2005). What does it mean to be a nonprofit educational measurement organization in the 21st century?. En R. Bennett y M. von Davier (Eds.), *Advancing human assessment. The methodological, psychological and policy contributions of ETS* (pp. 1-15). Springer.
- Boone, W. y Noltemeyer, A. (2017). Rasch analysis: A primer for school psychology researchers and practitioners. *Cogent Education*, 4(14), 1-13. <https://doi.org/10.1080/2331186X.2017.1416898>
- Buendía, M. A. y Rivera, R. (2010). Modelo de selección para el ingreso a la Educación Superior: El caso de la UACH. *Revista de la Educación Superior*, 39(156), 55-72.
- Buntis, M., Buntis, K. y Eggert, F. (2017). Clarifying the concept of validity: From measurement to everyday language. *Theory & Psychology*, 27(5), 703-710. <https://doi.org/10.1177/0959354317702256>
- Centro Nacional de Evaluación para la Educación Superior (CENEVAL). (2020). *EXANI-II Admisión*. CENEVAL.
- Cizek, G. J. (2001). More unintended consequences of high-stakes testing. *Educational Measurement: Issues and Practice*, 20(4), 19-27. <https://doi.org/10.1111/j.1745-3992.2001.tb00072.x>
- Cook, D. A., Bordage, G. y Schmidt, H. G. (2008). Description, justification and clarification: A framework for classifying the purposes of research in medical education. *Medical Education*, 42(2), 128-133. <https://doi.org/10.1111/j.1365-2923.2007.02974.x>
- Dirección General de Administración Escolar (DGAE) UNAM. (2019). *Demanda e ingreso a la licenciatura*. Recuperado de [http://www.estadistica.unam.mx/series\\_inst/index.php](http://www.estadistica.unam.mx/series_inst/index.php)

- Dirección General de Administración Escolar (DGAE) UNAM. (2020). *Acerca de nosotros, quiénes somos y qué hacemos. DGAE, UNAM.CdMx*. Recuperado de [https://www.dgae.unam.mx/acerca\\_nosotros.html](https://www.dgae.unam.mx/acerca_nosotros.html).
- Dirección General de Planeación (DGPL) UNAM. (2020). *Agenda Estadística 2020 UNAM*. Recuperado de: <http://www.estadistica.unam.mx/agenda.php>.
- Dorans, N. J. y Holland, P. W. (1992). *DIF detection and description: Mantel-Haenszel and standardization. ETS Research Report Series*, 1992, 1-40. <https://doi.org/10.1002/j.2333-8504.1992.tb01440.x>
- Frey, M. C. y Detterman, D. K. (2003). Scholastic assessment or G? The relationship between the Scholastic Assessment Test and general cognitive ability. *Psychological Science*, 15(6), 373-378. <https://doi.org/10.1111/j.0956-7976.2004.00687.x>
- Gago, A. (2000). El CENEVAL y la evaluación externa de la educación en México. *Revista Electrónica de Investigación Educativa*, 2(2).
- García-Medina, A. M., Martínez-Rizo, F. y Cordero Arroyo, G. (2016). Análisis del funcionamiento diferencial de los ítems del Excale de matemáticas para tercero de secundaria. *Revista Mexicana de Investigación Educativa*, 21(71), 1191-1220.
- Graue, E. (2018). *Acuerdo que reorganiza las funciones y estructura de la Secretaría General de la Universidad Nacional Autónoma de México*. Gaceta UNAM.
- Gregory, J. C. (2016). Validating test score meaning and defending test score use: different aims, different methods. *Assessment in Education: Principles, Policy & Practice*, 23(2), 212-225. <https://doi.org/10.1080/0969594X.2015.1063479>
- Guzmán, C., y Serrano, O. (2011). Las puertas del ingreso a la educación superior: el caso del concurso de selección a la licenciatura de la UNAM. *Revista de la Educación Superior*, 40(157), 31-53.
- Haladyna, T. M., Downing, S. M. y Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, 15(3), 309-334. [https://doi.org/10.1207/S15324818AME1503\\_5](https://doi.org/10.1207/S15324818AME1503_5)
- Holland, P. y Weiner, H. (1993). *Differential Item Functioning*. Laurence Erlbaum Associates.
- Juarros, M. (2006). ¿Educación superior como derecho o como privilegio?: Las políticas de admisión a la universidad en el contexto de los países de la región. *Andamios*, 3(5), 69-90.
- Kane, M. T. (2016). Explicating validity. *Assessment in Education: Principles, Policy & Practice*, 23(2), 198-211. <https://doi.org/10.1080/0969594X.2015.1060192>
- Kane, M. y Bridgeman, B. (2017). Research on validity theory and practice at ETS. En R. Bennett y M. von Davier (Eds.), *Advancing Human Assessment. Methodology of Educational Measurement and Assessment* (pp. 489-551). Springer.
- Lane, S., Raymond, M. R., Haladyna, T. M. y Downing, S. M. (2016). Test development process. En S. Lane, M. R. Raymond y T. M. Haladyna (Eds.), *Handbook of test development* (pp. 3-18). Routledge.
- Linacre J. M. y Wright, B. D. (1989). Mantel-Haenszel DIF and PROX are equivalent! *Rasch Measurement Transactions*, 3(2), 52-53.
- Manzi, J., Bosch, A., Bravo, D., del Pino, G., Donoso, G. y Pizarro, R. (2010). Validez diferencial y sesgo en la predictividad de las pruebas de admisión a las universidades chilenas. *Revista Iberoamericana de Evaluación Educativa*, 3(2), 30-48.
- Martínez-González, A., Sánchez-Mendiola, M., Manzano-Patiño, A., García-Minjares, M., Herrera-Penilla, C. y Buzo-Casanova, E. (2018). Grado de conocimientos de los estudiantes

- al ingreso a la licenciatura y su asociación con el desempeño escolar y la eficiencia terminal. Modelo multivariado. *Revista de la Educación Superior*, 47(188), 57-85.
- Martínez-Rizo, F. (2001). Evaluación educativa y pruebas estandarizadas. Elementos para enriquecer el debate. *Revista de la Educación Superior*, 30(120), 71-85.
- Martínez-Rizo, F. (2016). Impacto de las pruebas en gran escala en contextos de débil tradición técnica: Experiencia de México y el Grupo Iberoamericano de PISA. *RELIEVE*, 22(1), MO. <http://dx.doi.org/10.7203/relieve.22.1.8244>
- Mendoza, A. (2015). La validez en los exámenes de alto impacto: Un enfoque desde la lógica argumentativa. *Perfiles Educativos*, 37(149), 169-186.
- Mislevy, R. J. (2016). How developments in psychology and technology challenge validity argumentation. *Journal of Educational Measurement*, 53(3), 265-292. <https://doi.org/10.1111/jedm.12117>
- OCDE (2018). "How do admission systems affect enrolment in public tertiary education?" *Education Indicators in Focus*. Recuperado de <https://www.oecd-ilibrary.org/deliver/41bf120b-en.pdf?itemId=%2Fcontent%2Fpaper%2F41bf120b-en&mimeType=pdf> <https://doi.org/10.1787/41bf120b-en>
- Ordorika, I. Rodríguez, R. A. y Montes de Oca, M. M. (2013). Estudio Comparativo de Universidades Mexicanas. Fichas Institucionales 2007-2011. En DGEI-UNAM (Eds.), *Cuadernos de Trabajo de la Dirección General de Evaluación Institucional* (pp. 227-230). DGEI-UNAM.
- Patterson, F., Roberts, C., Hanson, M. D., Hampe, W., Eva, K., Ponnampuruma, G., et al. (2018). 2018 Ottawa consensus statement: Selection and recruitment to the healthcare professions. *Medical Teaching*, 40(11), 1091-1101. <https://doi.org/10.1080/0142159X.2018.1498589>
- Raykov, T. y Marcoulides, G. A. (2016). On the relationship between classical test theory and item response theory: From one to the other and back. *Educational and Psychological Measurement*, 76(2), 325-338. <https://doi.org/10.1177/0013164415576958>
- Ringsted, C., Hodges, B. y Scherpbier, A. (2011). The research compass: An introduction to research in medical education: AMEE Guide n° 56. *Medical Teaching*, 33(9), 695-709. <https://doi.org/10.3109/0142159X.2011.595436>
- Sánchez Mendiola, M., Delgado Maldonado, L., Flores Hernández, F., Leenen, I. y Martínez González, A. (2015). Evaluación del aprendizaje. En M. Sánchez Mendiola, A. Lifshitz Guinzberg, P. Vilar Puig, A. Martínez González, M. Varela Ruiz, M. y E. Graue Wiechers, (Eds.), *Educación Médica: Teoría y Práctica* (pp. 89-95). Elsevier.
- Sánchez-Mendiola, M. y Delgado-Maldonado, L. (2017). Exámenes de alto impacto: Implicaciones educativas. *Investigación Educativa Médica*, 6(21), 52-62. <https://doi.org/10.1016/j.riem.2016.12.001>
- Schuwirth, L., Colliver, J., Gruppen, L., Kreiter, C., Mennin, S., Onishi, H., et al. (2011). Research in assessment: Consensus statement and recommendations from the Ottawa 2010 Conference. *Medical Teaching*, 33(3), 224-233. <https://doi.org/10.3109/0142159X.2011.551558>
- Shepard, L. (2016). Evaluating test validity: Reprise and progress. *Assessment in Education: Principles, Policy & Practice*, 23(2), 268-280. <https://doi.org/10.1080/0969594X.2016.1141168>
- Sigal, V. y Dávila, M. (2004). La cuestión de la admisión a los estudios universitarios en Argentina. En O. Barsky, V. Sigal y M. Dávila (Eds.), *Los desafíos de la universidad argentina* (pp. 205-222). Siglo XXI Editores.

- Sireci, S. G. (2016). On the validity of useless tests. *Assessment in Education: Principles, Policy & Practice*, 23(2), 226-235. <https://doi.org/10.1080/0969594X.2015.1072084>
- Trost, G. (1993). Principios y prácticas en la selección para la admisión a la educación superior. *Revista de la Educación Superior*, 22(85), 1-10.
- UNAM. (1997). *Reglamento General de Inscripciones. Universidad Nacional Autónoma de México*. Recuperado de <https://www.dgae-siae.unam.mx/acerca/normatividad.html#leg-3>.
- Walker, C. (2011). What's the DIF? Why differential item functioning analyses are an important part of instrument development and validation. *Journal of Psychoeducational Assessment*, 29(4), 364-376. <https://doi.org/10.1177/0734282911406666>
- Yavuz, S., Dogan, N., Hambleton, R. K. y Yurtcu, M. (2018). The comparison of differential item functioning predicted through experts and statistical techniques. *Cypriot Journal of Educational Science*, 13(2), 375-384. <https://doi.org/10.18844/cjes.v13i2.2427>
- Young, M., St-Onge, C., Xiao, J., Vachon Lachiver, E. y Torabi, N. (2018). Characterizing the literature on validity and assessment in medical education: a bibliometric study. *Perspectives on Medical Education*, 7(3), 182-191. <https://doi.org/10.1007/s40037-018-0433-x>
- Zieky, M. (1993). DIF statistics in test development. En P. W. Holland y H. Wainer (Eds), *Differential item functioning* (pp. 337-347). Erlbaum.
- Zwick, R. (2006). Higher Education Admissions Testing, En R. Brennan (Ed.), *Educational Measurement* (pp. 647-679). National Council on Measurement in Education Greenwood Press.

## Breve Cv de los autores

### Melchor Sánchez Mendiola

Médico pediatra por la Universidad del Ejército y Fuerza Aérea, México; Maestro en Educación en Profesiones de la Salud por la Universidad de Illinois en Chicago, EUA; Doctor en Ciencias de la Educación por la UNAM. Profesor de Carrera Titular C de Tiempo Completo Definitivo, División de Estudios de Posgrado, Facultad de Medicina, Universidad Nacional Autónoma de México (UNAM). Participa en proyectos de evaluación educativa y educación en profesiones de la salud. ORCID ID: <https://orcid.org/0000-0002-9664-3208>. Email: [melchorsm@unam.mx](mailto:melchorsm@unam.mx)

### Manuel García Minjares

Actuario con estudios de Maestría en Estadística e Investigación de Operaciones por la Universidad Nacional Autónoma de México (UNAM). Profesor de Estadística, Probabilidad, Matemáticas y Operaciones en los sistemas escolarizados y a distancia de la Facultad de Contaduría y Administración de la UNAM. Actualmente es Jefe de la Unidad de Estadística y Análisis de Datos de la Dirección de Evaluación Educativa de la UNAM. ORCID ID: <https://orcid.org/0000-0002-9535-5917>. Email: [mminjares@unam.mx](mailto:mminjares@unam.mx)

### Adrián Martínez González

Médico Cirujano por la Universidad Nacional Autónoma de México. Doctor en Salud Pública y Medicina Preventiva por la Universidad Autónoma de Madrid. Profesor de Carrera Titular C Tiempo Completo Definitivo, Facultad de Medicina, UNAM. Miembro de la Academia Nacional de Medicina de México y del Sistema Nacional de Investigadores. Actualmente Director de Evaluación Educativa de la UNAM. Participa en proyectos de

evaluación educativa y educación en profesiones de la salud. ORCID ID: <https://orcid.org/0000-0002-5021-9639>. Email: [adrianmartinez38@gmail.com](mailto:adrianmartinez38@gmail.com)

**Enrique Buzo Casanova**

Licenciatura en Psicología por la Universidad Nacional Autónoma de México (UNAM). Especialidad en psicoterapia de corte psicoanalítico por la UNAM. Profesor de Asignatura de la Facultad de Psicología de la UNAM. Miembro del Colegio Nacional de Psicólogos de México. Actualmente Subdirector de Evaluación de Bachillerato y Licenciatura, en la Dirección de Evaluación Educativa de la UNAM. Participa en proyectos de evaluación educativa y educación de bachillerato y licenciatura. ORCID ID: <https://orcid.org/0000-0001-7490-7826>. Email: [erbuzo@unam.mx](mailto:erbuzo@unam.mx)