



Universidad de Valladolid

Facultad de Educación y Trabajo Social

Máster en Investigación Aplicada a la Educación

Trabajo Fin de Máster

**Puntos de corte en pruebas referidas a criterios:
Análisis comparativo de estrategias para la evaluación
de competencias digitales**

Presentado por:

Melissa Villalobos García

Directores:

Dra. Rocío Anguita Martínez

Dr. José María Marbán Prieto

Valladolid, junio 2018

Dedicatoria

A mi familia:

Por el apoyo incondicional de siempre y, especialmente en este año, por la importante compañía a pesar de los 8500kms de distancia. ¡Los amo!

Reconocimientos

El desarrollo de un proyecto de investigación usualmente está vinculado con importantes procesos de colaboración que, finalmente, nos permiten llevar el objetivo a buen puerto. En este caso, el desarrollo de mi Trabajo Fin de Máster, estuvo acompañado de importantes colaboraciones institucionales, académicas y personales que quiero agradecer.

En primer lugar, al Programa de becas para Iberoamérica + Asia de la Universidad de Valladolid y el Banco Santander de España, quienes vieron en mi perfil una persona con capacidad para aprovechar el apoyo económico facilitado para completar el programa de estudios de manera responsable y satisfactoria. En especial, agradezco a Esmeralda Lorenzo y al personal del Servicio de Relaciones Internacionales de la Universidad de Valladolid, quienes apoyaron el proceso formal para obtener la beca.

En el mismo nivel de prioridad, debo agradecer a la Fundación Omar Dengo de Costa Rica por el gran apoyo que me brindaron desde que empezó este proceso de realizar mis estudios de posgrado en el extranjero. Tanto al personal de Dirección Ejecutiva y de Recursos Humanos como a mis jefaturas Magaly Zúñiga y Melania Brenes, agradezco la confianza diaria en mi trabajo y el respaldo brindado para hacer que las puertas de la institución quedaran abiertas para mí al finalizar este máster.

En segundo lugar, a mi equipo de tutores con quienes empecé a gestionar esta idea de proyecto desde el inicio del curso académico y quienes depositaron su confianza en mis capacidades. A Rocío, un agradecimiento especial por ocuparse no sólo de los aspectos vinculados al TFM sino también por estar pendiente de mi proceso de adaptación al nuevo país. A José María, el agradecimiento especial es por ayudarme a ampliar mi horizonte y por motivarme a explorar nuevos caminos dentro nuestro querido campo de la investigación.

A Benito Arias y otros profesores del máster que de alguna manera aportaron alguna idea o conocimiento que me ayudaron a concretar el TFM. Especialmente a Jairo Rodríguez del programa del Doctorado, quien dedicó numerosas horas a acompañar mis avances en todo lo relacionado con la estrategia de *Science Mapping*. Finalmente, a mis compañeros de trabajo de la Unidad de Evaluación de la FOD que estuvieron vinculados con los avances de este proyecto y mis compañeros y compañeras del Máster que contribuyeron con mucho ánimo y compañía a la consecución de esta meta académica. ¡A todos y todas, muchas gracias!

Resumen

A través de una estrategia de revisión creativa de la literatura conocida como *Science Mapping*, se lleva a cabo una exploración del campo de conocimiento del establecimiento de puntos de corte en pruebas de evaluación referidas a criterios. Partiendo de un conjunto de 601 artículos científicos, se explora y caracteriza la configuración intelectual del campo de conocimiento y se identifican 17 estrategias de definición de puntos de corte.

Se realiza una categorización de la tipología de las estrategias identificadas, así como un análisis de su procedimiento, sus fortalezas y debilidades. Este análisis se toma como base para desarrollar una valoración de la pertinencia de cada estrategia según el contexto de la evaluación de competencias digitales en estudiantes beneficiarios del Programa Nacional de Informática Educativa, del Ministerio de Educación Pública y la FOD en Costa Rica.

Tomando como punto de partida los requerimientos metodológicos propios de dicho contexto, se define la estrategia *Bookmark* como la más oportuna de implementar. En función de ello, se define una propuesta de implementación en cuatro fases para llevarse a cabo desde la Unidad de Evaluación de la FOD, durante el segundo semestre del año 2018.

En términos generales, se considera que la FOD cuenta con importantes avances metodológicos que son oportunos para implementar de manera satisfactoria dicha estrategia de definición de puntos de corte. Por tanto, se pretende que el presente TFM contribuya a la orientación teórica y procedimental para tal proceso.

Palabras clave: puntos de corte en pruebas de evaluación, evaluación referidas a criterios, niveles de logro de estudiantes, evaluación de competencias digitales, bookmark method, revisión creativa de la literatura, science mapping.

Abstract

Through a strategy of creative review of the literature known as Science Mapping, an exploration of the establishment of cut points in criteria referred evaluation tests field is carried out. Based on a set of 601 scientific articles, the intellectual configuration of the field of knowledge is explored and characterized, and 17 strategies for definition of cut points are identified.

A categorization of the typology of the identified strategies is carried out, as well as an analysis of its procedure, its strengths and weaknesses. This analysis is taken as a basis to develop an assessment of the relevance of each strategy according to the context of the evaluation of digital skills in students who are beneficiaries of the National Program for Educational Information Technology, the Ministry of Public Education and the FOD (PRONIE MEP-FOD), in Costa Rica.

Taking as a starting point the methodological requirements of this context, the Bookmark method is defined as the most opportune to implement there. Based on this, an implementation proposal is defined in four phases to be carried out from the Evaluation Unit of the FOD, during the second semester of 2018.

In general terms, it is considered that the FOD has important methodological advances that are opportune to satisfactorily implement this strategy of definition of cut-off points in their evaluation tests. Therefore, it is intended that the present study contribute to the theoretical and procedural guidance for such a process.

Keywords: cut scores, evaluation tests, criteria referred evaluation, student achievement, digital competences evaluation, bookmark method, creative review of literature, science mapping.

Índice general

Dedicatoria	iii
Reconocimientos	v
Resumen.....	vii
Índice de Tablas	xii
Índice de Figuras.....	xiii
I. Introducción	1
A. Planteamiento del problema	3
B. Propósito del estudio	5
<i>Objetivo General</i>	5
<i>Objetivos Específicos</i>	5
C. Contextualización sobre el sistema educativo de Costa Rica.....	6
<i>El Programa Nacional de Informática Educativa del MEP y la FOD</i>	10
<i>Laboratorios de Informática Educativa del PRONIE MEP-FOD</i>	14
II. Marco de referencia.....	17
A. Marco conceptual.....	17
<i>Competencias digitales en el ámbito educativo</i>	17
<i>Estándares de desempeño de estudiantes en el aprendizaje con tecnologías digitales del PRONIE MEP-FOD</i>	21
<i>Evaluación referida a criterios y establecimiento de puntos de corte</i>	25
<i>Science Mapping como estrategia de revisión creativa de la literatura</i>	31
B. Antecedentes del estudio.....	33
III. Marco metodológico	39
A. Tipo de estudio	39
B. Muestra	41
C. Procedimiento de recolección y análisis de los datos.....	44

1.	<i>Fase preparatoria</i>	44
2.	<i>Fase descriptiva</i>	45
3.	<i>Fase analítica</i>	45
4.	<i>Fase propositiva</i>	46
IV.	Resultados	47
A.	Configuración intelectual del campo de conocimiento	47
	<i>Descripción del conjunto de datos</i>	47
	<i>Análisis de redes y mapas de tendencias del conjunto de datos</i>	51
B.	Análisis de las estrategias identificadas según pertinencia para la FOD	59
	<i>Caracterización general de las estrategias identificadas</i>	61
	Valoración de fortalezas y debilidades de las estrategias	64
C.	Propuesta de implementación en el contexto evaluativo de la FOD	70
	Fase 1. Revisión y conceptualización de los niveles de logro	71
	Fase 2. Selección y capacitación del grupo de jueces	74
	Fase 3. Obtención de los puntos de corte para la prueba	76
	Fase 4. Interpretación de resultados de la evaluación.....	79
V.	Conclusiones	81
A.	Sobre los resultados del estudio	81
B.	Alcances y limitaciones	83
C.	Recomendaciones y líneas de investigación futura.....	85
VI.	Referencias	87
VII.	Anexos	95
A.	Anexo 1. Fases de evaluación de competencias digitales en la FOD	95
B.	Anexo 2. Línea del tiempo de la conformación del PRONIE MEP-FOD.....	96
C.	Anexo 3. Cronograma de trabajo para el desarrollo del TFM.....	97
D.	Anexo 4. Script ejecutado en R para el análisis bibliométrico	98

E.	Anexo 5. Análisis descriptivo del conjunto de datos	99
F.	Anexo 6. Fichas síntesis de las estrategias analizadas.....	103
	<i>Estrategias centradas en el test</i>	103
	<i>Estrategias centradas en el examinado</i>	106
	<i>Estrategias de compromiso</i>	109
G.	Anexo 7. Síntesis de sesión de validación de análisis con equipo FOD	111
H.	Anexo 8. Ejemplo de plantilla para formulario de respuesta de los jueces.....	112
I.	Anexo 9. Propuesta de agenda de trabajo en sesión con los jueces.....	113
J.	Anexo 10. Escala de valoración de la sesión por parte de los jueces.....	115

Índice de Tablas

Tabla 1.	Conceptualización del modelo de Estándares de desempeño de estudiantes en el aprendizaje con tecnologías digitales del PRONIE MEP-FOD	22
Tabla 2.	Diferencias entre pruebas referidas a normas y pruebas referidas a criterios	26
Tabla 3.	Tipos de información utilizada para dar realimentación a los jueces	29
Tabla 4.	Criterios de rigurosidad científica tomados en cuenta en el estudio	40
Tabla 5.	Principales componentes del <i>ISI Web of Science</i> que integran el universo muestral del TFM	41
Tabla 6.	Procedimiento para la primera búsqueda de información en la WOS	42
Tabla 7.	Procedimiento para la segunda búsqueda de información en la WOS	43
Tabla 8.	Información principal del conjunto de datos	49
Tabla 9.	Otras referencias frecuentemente citadas dentro del conjunto de datos	51
Tabla 10.	Criterios de comparación de estrategias contemplados en el análisis	60
Tabla 11.	Distribución de estrategias según requerimiento de datos empíricos y tipo de punto de corte que genera su implementación	62
Tabla 12.	Tipos de ítem de las estrategias prioritarias para el contexto de la FOD	63
Tabla 13.	Modelo de categorización de la pertinencia de las estrategias según el contexto de evaluación de la FOD	64
Tabla 14.	Síntesis de valoraciones para las estrategias categorizadas como Pertinencia Baja	66
Tabla 15.	Síntesis de valoraciones para las estrategias categorizadas como Pertinencia Media	67
Tabla 16.	Rondas de valoración a realizar durante la implementación de la estrategia <i>Bookmark</i>	77

Índice de Figuras

Figura 1.	Síntesis conceptual sobre la que se fundamenta el ámbito disciplinar del estudio	4
Figura 2.	Evolución de la cobertura y estudiantes beneficiados por el PRONIE MEP-FOD en el periodo 2014-2017	11
Figura 3.	Evolución de los aportes del PRONIE MEP-FOD al cierre de la brecha digital en Costa Rica	12
Figura 4.	Nivel de logro de los estudiantes en las dimensiones de los Estándares de desempeño evaluados en el 2016	15
Figura 5.	Modelo de dimensiones para la sistematización del concepto de competencia digital	18
Figura 6.	Dimensiones del Marco DIGCOMP para la conceptualización de la Competencia Digital a nivel europeo	20
Figura 7.	Dimensiones y propiedades de los Estándares de desempeño del PRONIE MEP-FOD	22
Figura 8.	Detalle de la estructura conceptual para la evaluación de los Estándares de Desempeño del PRONIE MEP-FOD	23
Figura 9.	Distribución de los Estándares de desempeño e hitos en las dimensiones a estudiar en el 2018	24
Figura 10.	Síntesis de aspectos a considerar para elegir una estrategia de definición de puntos de corte	30
Figura 11.	Diagrama de flujo para la implementación de la estrategia de <i>Science Mapping</i>	32
Figura 12.	Tabla de resultados de puntos de corte establecidos mediante el método Angoff	35
Figura 13.	Tabla de resultados de puntos de corte establecidos mediante el método Bookmark	35
Figura 14.	Comparación entre las metodologías <i>Angoff</i> , <i>BoW</i> (<i>Cuerpo de Trabajo</i>), <i>Juicio analítico</i> y <i>Bookmark</i> siguiendo criterios de Berk (1986)	38
Figura 15.	Síntesis del procedimiento de recolección y análisis de datos llevado a cabo para el desarrollo del TFM	44

Figura 16.	Porcentaje de aporte por país al campo de conocimiento estudiado	49
Figura 17.	Análisis de co-ocurrencia de palabras clave definidas por los autores [<i>authors keywords</i>] y las asignadas por la base de datos [<i>keywordsplus</i>]	52
Figura 18.	Análisis de co-ocurrencia de términos del título y resumen de los artículos	54
Figura 19.	Evolución temporal de los principales términos de las publicaciones en el campo de conocimiento	55
Figura 20.	Análisis de citación entre países dentro del conjunto de datos	58
Figura 21.	Evolución temporal de la citación entre países dentro del conjunto de datos	57
Figura 22.	Clasificación de las estrategias de definición de puntos de corte para pruebas de evaluación referidas a criterios, identificadas en la literatura	61
Figura 23.	Categorización las estrategias según pertinencia para la FOD	65
Figura 24.	Síntesis de ventajas y desventajas identificadas en la estrategia <i>Angoff V1</i>	68
Figura 25.	Síntesis de ventajas y desventajas identificadas en la estrategia <i>Bookmark</i>	68
Figura 26.	Ejemplo de OIB en la estrategia <i>Bookmark</i>	70
Figura 27.	Fases de implementación de la estrategia <i>Bookmark</i> para definir los puntos de corte en la prueba de evaluación de la FOD.	71
Figura 28.	Categorización de los niveles de habilidad utilizados en las pruebas de la FOD, según quintiles	72
Figura 29.	Síntesis de pasos asociados a la Fase de revisión y conceptualización de los niveles de logro.	73
Figura 30.	Síntesis de pasos asociados a la Fase de selección y capacitación de los jueces.	76
Figura 31.	Síntesis de pasos asociados a la Fase de obtención de los puntos de corte.	78
Figura 32.	Variables independientes a considerar en la profundización de factores asociados a los niveles de logro de los estudiantes en la prueba.	79

I. Introducción

La Fundación Omar Dengo (en adelante FOD), como institución pionera en el sistema educativo costarricense en el desarrollo de competencias digitales ha realizado en los últimos años, desde el Área de Investigación y Evaluación, evaluaciones formativas y de resultados de sus propuestas educativas y de las del Programa Nacional de Informática Educativa del Ministerio de Educación Pública de Costa Rica y la FOD (en adelante PRONIE MEP-FOD).

Desde el año 2011 la FOD ha trabajado en el desarrollo de un proyecto de evaluación cuyo objetivo es “Evidenciar el nivel de logro de los Estándares de Desempeño de estudiantes en el aprendizaje con tecnologías digitales en los estudiantes del PRONIE MEP-FOD” (FOD, 2014a). En el marco de este proyecto, se define un conjunto Estándares de Desempeño para delimitar qué se espera que los y las estudiantes sepan y puedan hacer con las tecnologías digitales en los Laboratorios de Informática Educativa del PRONIE MEP-FOD (en adelante LIE).

En términos de evaluación, se han llevado a cabo diferentes fases para evaluar de manera periódica los resultados de los y las estudiantes de Primaria y Secundaria en los que se cuenta con mayor cantidad de población beneficiaria (FOD 2014a, 2015 y 2017a). En estas evaluaciones se han obtenido resultados en los que ha sido posible comparar el desempeño de cada estudiante con su propio grupo, es decir, se han realizado análisis más orientados a la naturaleza de las pruebas de evaluación referidas a la norma (Véase la síntesis de las fases de evaluación del PRONIE MEP-FOD en el **Anexo 1**).

Con el objetivo de generar resultados que aporten información más específica sobre el nivel de logro de la población estudiantil, recientemente se ha trabajado en el establecimiento de unos puntos de corte mediante métodos de base empírica, clasificando a priori los ítems de las pruebas según un nivel de dificultad para comparar posteriormente los resultados de los estudiantes con éstos. Este paso ha permitido ofrecer resultados más cercanos a la lógica de las pruebas referidas a criterios, en las que se busca proporcionar resultados sobre el nivel de conocimiento o de dominio que tiene el estudiante en relación con una competencia.

La experiencia ha sido útil para evidenciar la necesidad de explorar y profundizar en posibles técnicas psicométricas para establecer puntos de corte que permitan definir los niveles de habilidad de los y las estudiantes. Ante este escenario, se visualizó como una

oportunidad importante el aprovechar los conocimientos adquiridos en el Máster de Investigación Aplicada a la Educación de la Universidad de Valladolid, realizando un Trabajo Fin de Máster (en adelante TFM) que permitiese contribuir a la mejora de la calidad de estas pruebas de evaluación. Partiendo de ello, se propone llevar a cabo un análisis comparativo de las principales estrategias psicométricas útiles para definir puntos de corte en una prueba de evaluación referida a criterios, como la descrita anteriormente.

Desde la literatura se plantea que no es posible definir a priori una estrategia de definición de puntos de corte como la más adecuada (Cizek y Bunch, 2007), sino que su selección debe hacerse en función del objetivo evaluativo, que delimita el tipo de competencia que se desea evaluar, el tipo de prueba que se construye y el tipo de información que se quiere obtener, entre otros aspectos. Partiendo de esto, la expectativa del TFM es generar un análisis que permita orientar la elección de una estrategia de definición de puntos de corte pertinente para las pruebas de evaluación que se realizan en la FOD.

Con el fin último de contribuir a la mejora de la calidad de las evaluaciones llevadas a cabo en el ámbito educativo costarricense, es posible derivar el siguiente objeto de estudio como eje para el desarrollo del presente TFM: *Criterios de selección de metodologías más pertinentes para el establecimiento de puntos de corte en pruebas de evaluación referidas a criterios que evalúan desarrollo de competencias digitales, en estudiantes de Primaria de Costa Rica.*

Partiendo del hecho de que la producción científica ha experimentado un crecimiento y evolución exponencial en los últimos 15 años (Scharnhorst, Börner, & Van den Besselaar, 2012), el análisis cuantitativo de ésta requiere del uso de técnicas de análisis que permitan evidenciar de forma eficaz las dinámicas, las estructuras y la evolución del campo de conocimiento al que responde. De esta manera, considerando las oportunidades que la web 2.0 ofrece en la actualidad al campo de la investigación, la aproximación a partir de la cual se aborda el objeto de estudio del presente TFM corresponde a una estrategia de revisión creativa de la literatura definida como *Science Mapping*.

Dicha estrategia se considera oportuna para explorar a profundidad la configuración intelectual de un campo de conocimiento (Rodríguez-Bolívar, Alcaide-Muñoz & Cobo, 2018), por lo que resulta útil para conducir el análisis propuesto. Derivado de esto, el estudio se plantea como una oportunidad para ampliar las evidencias empíricas en torno al apoyo que las tecnologías digitales pueden brindar actualmente al campo de la investigación en educación.

A. Planteamiento del problema

La presente investigación se enmarca dentro del área temática del uso de las Tecnologías Digitales en la Educación, específicamente en lo que se relaciona con el desarrollo de competencias digitales en estudiantes mediante su uso. El ámbito disciplinar dentro del cual se inserta corresponde a la Psicometría, concretamente en el campo del desarrollo de pruebas referidas a criterios, siendo el foco principal el que se ocupa de las técnicas de establecimiento de puntos de corte.

En la **Figura 1** se muestra una síntesis de la articulación conceptual que fundamenta este estudio. Asimismo, es importante explicitar algunas de las principales premisas conceptuales de las cuales parte:

- ✓ El uso pedagógico de las tecnologías digitales en la educación favorece el desarrollo de habilidades cognitivas y competencias digitales en la población de Primaria.
- ✓ En Costa Rica, las habilidades que se busca desarrollar en la población estudiantil mediante el uso de tecnologías digitales se organizan en un marco conceptual denominado *Estándares de desempeño de estudiantes en el aprendizaje con tecnologías digitales*, los cuales se organizan en tres dimensiones: 1) Resolución de problemas e Investigación; 2) Productividad; y 3) Ciudadanía y comunicación (Véase el detalle en Marco conceptual).
- ✓ La Psicometría ofrece múltiples estrategias de definición de puntos de corte en pruebas de evaluación referidas a criterios que permiten aproximar de mejor manera el nivel de competencia o logro de los y las estudiantes.

El contexto descrito con anterioridad da pie a la formulación de tres interrogantes:

1. ¿Cuáles son las principales tendencias que caracterizan la configuración intelectual del campo de conocimiento sobre la definición de puntos de corte en pruebas de evaluación referidas a criterios?
2. ¿Cuáles son las fortalezas y debilidades de las principales metodologías de definición de puntos de corte en pruebas de evaluación referidas a criterios según su pertinencia para el ámbito de la educación?
3. ¿Qué estrategias de definición de puntos de corte en pruebas referidas a criterios evidencian ser más pertinentes para las evaluaciones de competencias digitales en estudiantes en el contexto educativo costarricense?

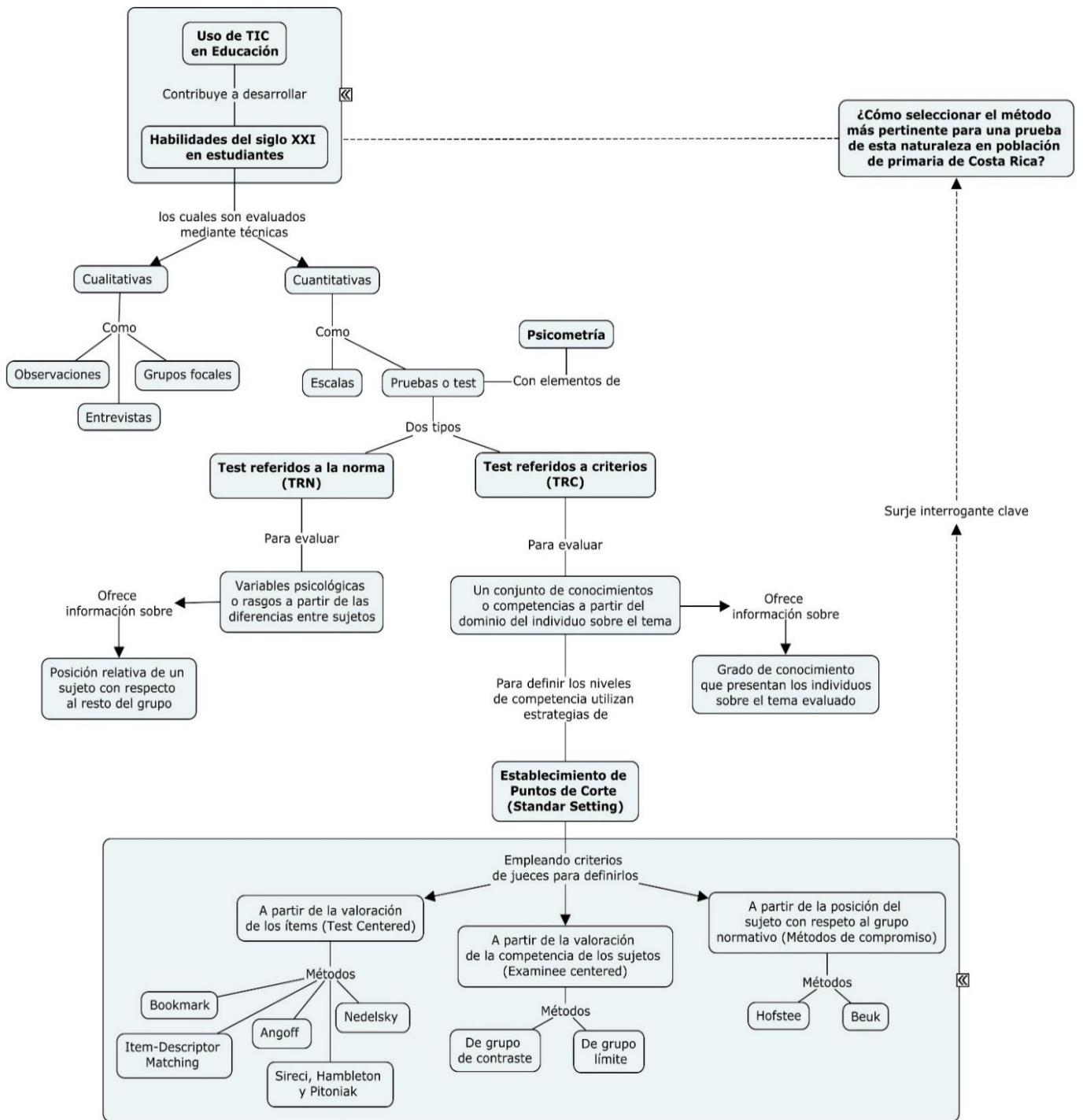


Figura 1. Síntesis conceptual sobre la que se fundamenta el ámbito disciplinar del estudio. (Fuente: Elaboración propia a partir de la revisión de la literatura y los intereses de investigación).

B. Propósito del estudio

Como se detalla en FOD (2018), existen diferentes estrategias metodológicas para definir puntos de corte en pruebas de evaluación referidas a criterios, cuyas características funcionan mejor según las características y los requerimientos de la evaluación. Es por esto, que en el contexto de la evaluación del logro de la población del PRONIE MEP-FOD “se necesita un estudio documental previo que detalle las metodologías existentes, con el objetivo de seleccionar la opción que mejor se adapte a los requerimientos de la investigación” (FOD, 2018, p. 32).

Esta necesidad se toma como base para la formulación de los objetivos de investigación del presente TFM, con el fin de apoyar los procesos de evaluación que se realizan desde dicho contexto. De esta manera, se definen los siguientes objetivos de investigación:

Objetivo General

Comparar las principales estrategias psicométricas para establecer puntos de corte en pruebas referidas a criterios en cuanto a sus alcances y limitaciones para contribuir a la mejora de la calidad de las evaluaciones de competencias digitales de estudiantes en el ámbito educativo costarricense.

Objetivos Específicos

1. Explorar la configuración actual del campo de conocimiento de definición de puntos de corte en pruebas de evaluación referidas a criterios mediante una estrategia de revisión creativa de la literatura.
2. Establecer las fortalezas y debilidades de las principales metodologías para establecer puntos de corte en pruebas referidas a criterios para evaluar competencias digitales en estudiantes.
3. Derivar criterios para planificar una estrategia de establecimiento de puntos de corte en pruebas que evalúan competencias digitales en estudiantes según su pertinencia para el contexto educativo costarricense.

C. Contextualización sobre el sistema educativo de Costa Rica

Costa Rica asumió desde muy temprano en su historia de nación independiente, la decisión de convertir a la educación en un pilar de su desarrollo y el de su gente.

Así, en 1847 la declara constitucionalmente como universal, gratuita y costeadada por el Estado, adelantándose a países de mayor desarrollo social, económico y demográfico de la región (Muñoz et al., 2014, p. 9).

El sistema educativo costarricense está organizado y supervisado por el Ministerio de Educación Pública (en adelante MEP), ente encargado de ejecutar las acciones necesarias para cumplir con la política educativa que se haya fijado en cada periodo para el país. Actualmente, está en vigencia la *Política Educativa del Siglo XXI*, cuyos objetivos corresponden a (MEP, 2017):

- ✓ Cerrar las brechas existentes entre la calidad de la educación que reciben los y las estudiantes de las áreas urbanas y rurales, y eliminar la diferenciación entre las instituciones educativas de las áreas urbanas marginales y no marginales.
- ✓ Formar recursos humanos que eleven la competitividad del país, necesaria para triunfar en los mercados internacionales.
- ✓ Fortalecer valores fundamentales que se han ido perdiendo con el pasar del tiempo.
- ✓ Fortalecer la educación técnica y científica, a la par de la deportiva y la cultural; como forma de estimular el desarrollo integral de los estudiantes.
- ✓ Hacer conciencia en los individuos, acerca del compromiso que tienen con las futuras generaciones, procurando un desarrollo sostenible económico y social, en armonía con la naturaleza y el entorno en general.

El sistema educativo de Costa Rica se divide entre la educación pública, la educación privada y la privada-subvencionada, es decir, que recibe algún tipo de ayuda económica parcial por parte del Estado. Pese a que existe dicha división, los datos más recientes del MEP (2016) muestran que cerca del 89% de los centros educativos del país corresponde al sector público (con un total de 4516 centros educativos), frente a un porcentaje que ronda el 12% en el sector privado (608 centros educativos) y menos del 2% de educación privada

subvencionada (cerca de 74 centros educativos). Esta estructura educativa se divide, a su vez, en cuatro niveles:

1. **Preescolar:** incluye el grado *Interactivo* que atiende población entre los 3 y los 4 años de edad, así como el de *Transición*, que atiende a los niños y niñas con edades entre los 5 y los 6 años.
2. **Educación General Básica:** integrado por 11 grados obligatorios que se dividen en dos bloques:
 - ✓ *Primaria:* comprende de primer a sexto grado, abarcando población entre los 6 y los 12 años de edad. Se subdivide en *I ciclo*, que abarca primero, segundo y tercer grado; y en *II ciclo*, que abarca cuarto, quinto y sexto grado.
 - ✓ *Secundaria:* comprende de séptimo a noveno grado; abarcando población que va desde los 12 hasta los 15 años. También se conoce como *III ciclo*, siguiendo con la denominación anteriormente mencionada.
3. **Educación diversificada:** denominada también como *IV ciclo*, comprende desde décimo a undécimo grado (o duodécimo, dependiendo de la modalidad elegida). Abarca población de los 16 a los 17 o 18 años y el estudiante puede elegir entre tres ramas de educación: académica, artística o técnica.
4. **Educación superior o universitaria:** que abarca pregrados, grados y posgrados, abarcando población con edades variables.

Como se mencionó en la introducción de este apartado y de acuerdo con el Artículo 78 de la Constitución Política de Costa Rica, “La educación preescolar, general básica y diversificada son obligatorias y, en el sistema público, gratuitas y costeadas por la Nación” (Constitución Política de la República de Costa Rica, 1949, p. 18). Este beneficio se refleja en el hecho de que es uno de los países del istmo con mayor matrícula en Primaria y Secundaria, contando en el año 2016 con un total de 951.227 estudiantes matriculados de un total de 4.301.712 habitantes en el país, según los datos más recientes (MEP, 2016).

Vinculado con esto, según los datos del último Informe Estado de la Educación del Programa Estado de la Nación de Costa Rica (PEN, 2017), las tasas de cobertura del sistema educativo costarricense son cercanas a los promedios de América Latina y el Caribe, aunque siguen siendo inferiores a las de los países pertenecientes a la Organización

para la Cooperación y el Desarrollo Económicos (OCDE). Al analizar en detalle las cifras de cobertura neta¹ del sistema educativo costarricense se identifica que (PEN, 2017):

- ✓ En Preescolar las tasas de cobertura del nivel *Interactivo II* mostraron un crecimiento leve, pasando de 57,4% en el año 2013 a 63% en 2016. Sin embargo, en el caso *Transición* se mostró una tendencia decreciente pasando de un 95,4% en el año 2008 a un 84,5% en el 2015.
- ✓ En Primaria se identificó una tendencia igualmente decreciente tanto en *I* como en *II ciclo*. Concretamente, en el año 2016 la tasa neta fue de 93,1%, siendo menor a lo mostrado en el período 2005-2011, cuando fue superior al 97%.
- ✓ En el caso de la educación Secundaria o *III ciclo*, las tasas de cobertura mostraron una tendencia creciente. Específicamente, aumentaron de 64,8% en el año 2010 a 73,3% en 2016. Sin embargo, se han identificado dificultades para que los estudiantes continúen del *III* al *IV ciclo* (educación diversificada), en donde para el año 2016 se tuvo una tasa neta del 45,8%.
- ✓ En el sector de educación superior la cobertura de la población en edad de asistir a este nivel (grupo de 18 a 22 años) alcanzó el 48,5% para el año 2015.

En materia de inversión en educación actualmente el gobierno de Costa Rica destina un 7,86% de su Producto Interno Bruto (PIB), colocándose como el país líder en la región en cuanto a esta inversión. Según datos de la OCDE, en Latinoamérica el promedio de inversión en educación ronda el 5%, con lo cual Costa Rica supera a países como Chile, cuya inversión en educación es del 6,4% del PIB, a Argentina, que dispone del 5,4% y Colombia, que invierte cerca del 4,9% (Barquero, 07 de agosto, 2017).

Además de ser líder en la región en materia de inversión en educación, el sistema educativo de Costa Rica muestra características ventajosas en relación con otros países. Por ejemplo, partiendo del hecho de que la Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO, por sus siglas en inglés) considera un país libre de analfabetismo aquel cuyo porcentaje no supera el 5%, según el Banco Centroamericano de Integración Económica (BCIE): “Costa Rica es el único país de Centroamérica libre de

¹ La cobertura neta mide la proporción de niños en un rango determinado de edad, que están asistiendo a los centros educativos, respecto a la población total que en esa edad debería asistir.

analfabetismo, con solo 3,2% de su población en esta condición” (Garza, 11 de enero de 2016).

Desde el MEP se ha intentado brindar a la población estudiantil las condiciones adecuadas para que logren mantenerse dentro del sistema educativo a pesar de las desigualdades sociales. De esta manera, un niño o niña que proviene de un hogar con clima educativo bajo tiene una probabilidad del 76% de terminar la Primaria gracias a las condiciones que se intenta brindar desde dentro del sistema educativo (PEN, 2017). Sin embargo, los datos muestran que al llegar a la Secundaria, los y las estudiantes muestran mayores dificultades para finalizarla. Por ejemplo, quienes cursaron el undécimo grado durante el año 2016 representaban el 45,4% de los que iniciaron séptimo año en el 2012 (PEN, 2017).

Algunas de las condiciones que el MEP ha buscado ofrecer a la población estudiantil son, por ejemplo, la disponibilidad de materiales educativos y de recursos tecnológicos en las aulas; aspectos que han mostrado estar asociados de manera positiva con el rendimiento académico (PEN, 2017). Según dicho informe, algunos otros aspectos influyentes en el rendimiento académico son las expectativas de los padres y madres de familia sobre el éxito de sus hijos e hijas, el nivel socioeconómico de las familias y la asistencia y puntualidad de los y las docentes.

Por otra parte, el rendimiento de los y las estudiantes costarricenses, según parámetros internacionales como el del Programa Internacional para la Evaluación de Estudiantes (PISA), se encuentra por debajo de los resultados de países miembros de la OCDE y evidencian dificultades, principalmente en competencia matemática. Pese a estos resultados, se destaca como un aspecto positivo:

(...) que hay alumnos que viven en condiciones de desventaja socioeconómica y, pese a ello, logran un buen desempeño, gracias a factores como el acceso a las tecnologías de información y comunicación (TIC) y el conocimiento sobre temas ambientales (PEN, 2017, p. 50).

Con el objetivo de mejorar los resultados de aprendizaje de la población estudiantil, uno de los principales retos a los que se enfrenta el sistema educativo costarricense es el de articular de manera más contundente las directrices estipuladas en las políticas educativas (por ejemplo en materia de currículo, infraestructura y equidad) con las necesidades reales que tienen estudiantes y docentes y con lo que realmente ocurre en las aulas (PEN, 2017).

Se considera que estudiar los factores asociados a los resultados de los y las estudiantes en este tipo de pruebas es potencialmente útil para que el MEP pueda continuar desarrollando y fortaleciendo acciones que beneficien el logro educativo. Un ejemplo de ello es el hallazgo sobre mejor desempeño académico por parte la población estudiantil que tiene mayor acceso a recursos tecnológicos, al respecto de lo cual se hace mención al PRONIE MEP-FOD en el siguiente apartado.

El Programa Nacional de Informática Educativa del MEP y la FOD

El Programa Nacional de Informática Educativa o PRONIE se implementa desde el año 1988 en Costa Rica gracias a una alianza público-privada entre el Ministerio de Educación Pública y la Fundación Omar Dengo. Vinculado a la Política Nacional para la Niñez y la Adolescencia, este programa está amparado bajo la Ley N° 8207 "Declaración de Utilidad Pública del Programa de Informática Educativa" (SITEAL, 2017).

El PRONIE MEP-FOD está destinado a personas de todo el territorio nacional, incluyendo niños, niñas y jóvenes estudiantes, así como docentes. Desde su visión, mediante el uso de las tecnologías digitales en los procesos de aprendizaje se busca incrementar la calidad de la educación costarricense apoyando el desarrollo de capacidades en las personas (Muñoz et al., 2014). Según se reseña en Muñoz et al. (2014, p.9) desde finales de la década de los 80's el Programa nace con el objetivo de contribuir concretamente con:

- ✓ La incorporación plena de las personas a la economía nacional y la dinámica internacional.
- ✓ Fomentar un sistema educativo de calidad y a la altura de los tiempos.
- ✓ Fortalecer los mecanismos de participación ciudadana y promover la cohesión social al interior del país, contribuyendo a la reducción de brechas geográficas, socioeconómicas, educativas y tecnológicas.

Estos objetivos han marcado la toma de decisiones en torno a la implementación del Programa, con lo cual se han destacado algunas características distintivas como el haber centrado la propuesta en aprendizaje de la programación como estrategia para desarrollar capacidades en los y las estudiantes, el establecimiento de vínculos con el currículo

académico nacional y el haber iniciado los esfuerzos de manera progresiva, iniciando con la educación Primaria (Muñoz, et al., 2014).

En el **Anexo 2** es posible visualizar una síntesis de la línea del tiempo de la conformación del PRONIE MEP-FOD y sus principales logros. Cabe destacar que con el paso de los años éste ha ido aumentando sus alcances y, como se muestra en la **Figura 2**, actualmente cuenta con una cobertura del 87,6% de la educación pública diurna (FOD, 2017b). Esta cifra corresponde a cerca de 3.219 centros educativos y 652.433 estudiantes beneficiados, abarcando población estudiantil desde el grado de *Preescolar* hasta el *III Ciclo* de la Educación General Básica.

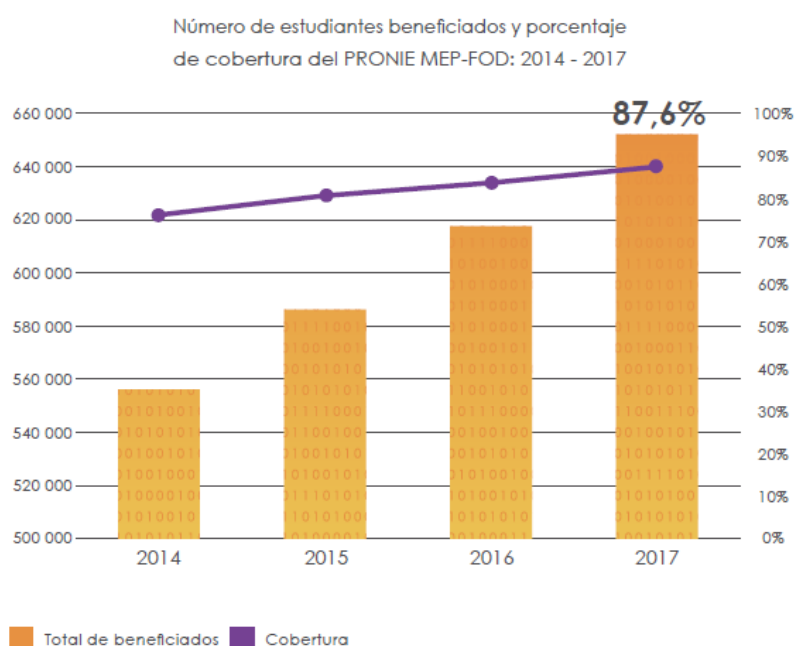


Figura 2. Evolución de la cobertura y estudiantes beneficiados por el PRONIE MEP-FOD en el periodo 2014-2017. (Fuente: FOD, 2017b, p. 20).

La puesta en práctica del modelo educativo del Programa se ha logrado a partir de la atención de una serie de condiciones y la provisión de recursos tecnológicos a los centros educativos, en los que se han incluido las computadoras como un objeto para construir y para pensar (Zúñiga, 2001). Un aspecto que ha diferenciado desde el inicio al PRONIE MEP-FOD de otros programas orientados a apoyar el aprendizaje mediante el uso de tecnologías digitales, es el siguiente:

Se tuvo claro que las TIC por sí mismas no harían la diferencia, y que su uso para el desarrollo de capacidades tendría resultado solo dentro de un marco pedagógico

centrado en la actividad de los estudiantes con las herramientas digitales, que les permitiera poner en práctica procesos de resolución de problemas y creación (Trilling, en Muñoz et al., 2014, p. 35).

Como se muestra en la **Figura 3**, el Programa contribuye al cierre de la brecha digital en el sistema educativo costarricense, aunque su principal enfoque está orientado a desarrollar habilidades en los y las estudiantes, así como en otras poblaciones beneficiarias. Partiendo de esto, la oferta educativa del PRONIE MEP-FOD ha estado orientada al desarrollo de habilidades, a la vez que ha velado por mantener sus propuestas vigentes y coherentes con los cambios y avances tecnológicos de las últimas décadas.

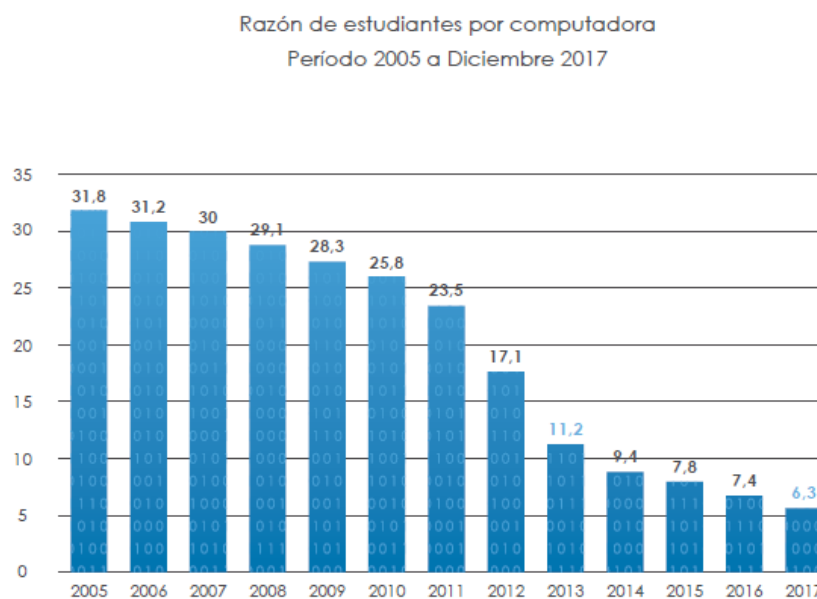


Figura 3. Evolución de los aportes del PRONIE MEP-FOD al cierre de la brecha digital en Costa Rica. (Fuente: FOD, 2017b, p. 26).

Actualmente la oferta educativa del PRONIE MEP-FOD se divide en cuatro grupos, dentro de los que se encuentran (FOD, 2017b):

- 1. Informática Educativa y Pensamiento Computacional.** Con propuestas como los LIE, en Primaria y Secundaria. Aunque sobre este espacio se profundizará más adelante, es importante mencionar que en él, se busca que los y las estudiantes aprendan a programar, diseñar videojuegos y hacer videos, entre otras actividades, como una vía para el desarrollo de destrezas cognitivas y sociales, como la resolución de problemas, la creatividad y la colaboración, entre otras (Muñoz, et al., 2014).

2. **Aprendizaje con Tecnologías Móviles.** Incluye propuestas como MoviLab en Primaria y Secundaria, las cuales promueven el aprovechamiento de las tecnologías digitales en aspectos curriculares concretos de las áreas de Ciencias, Matemáticas y Español. Otra de las propuestas enmarcadas dentro de este grupo, es la de Redes Móviles para el Aprendizaje o Rem@, desde la que se busca que la población estudiantil de Secundaria aproveche los recursos tecnológicos en aspectos personales, socio-productivos y sociales. Asimismo, las propuestas de Aprendizaje con Tecnologías Móviles en escuelas multigrado y en centros educativos indígenas se incluyen en el grupo como iniciativas que buscan el desarrollo de competencias para resolver problemas, investigar, producir, comunicarse con otras personas y llevar a cabo una práctica ciudadana responsable desde la Primaria (Muñoz et al. 2014).
3. **Capacidades emprendedoras y de Innovación.** Con propuestas como Robótica y Aprendizaje por Diseño. Estas propuestas iniciaron cerca de 1998 y, según mencionan Muñoz et al. (2014), fueron iniciativas pioneras para América Latina y el Caribe. En el caso de la robótica, basada igualmente en el aprendizaje de la programación, promueve un conjunto de desempeños y habilidades vinculados con la creatividad, el diseño, la construcción y la divulgación de creaciones propias de los y las estudiantes, representadas mediante prototipos resultantes de su imaginación. También destaca la propuesta de Labor@ que busca desarrollar, en los y las jóvenes de Secundaria, habilidades para incorporarse al mundo del trabajo y para generar emprendimientos.
4. **Desarrollo Profesional Docente.** Incluye el diseño y la ejecución de módulos de capacitación y formación para docentes, plataformas y recursos de apoyo, entre otros. Esta línea de trabajo ha sido uno de los pilares del Programa desde su origen, al considerarse como una prioridad el “fortalecer su papel [del cuerpo docente] como agentes dinamizadores para el cambio y el fortalecimiento de la calidad educativa en el sistema público” (Muñoz et al., 2014, p. 52).

En concordancia con el foco del presente trabajo de investigación, en el siguiente apartado se profundizará en la caracterización de la propuesta de LIE. Sin embargo, cabe finalizar esta descripción del Programa haciendo énfasis en que su estructura a nivel institucional incluye dentro de sus pilares el desarrollo de procesos de investigación y evaluación, con el objetivo de realimentar la práctica del Programa y contribuir a su enriquecimiento y avance (Muñoz et al., 2014).

Para ello, se han llevado a cabo procesos de evaluación externa y, constantemente, se desarrollan procesos de evaluación formativa interna. Estos procesos se visualizan como un mecanismo de rendición de cuentas a las autoridades educativas nacionales y, a la vez, como una forma de mantener las propuestas educativas vigentes mientras se monitorea el logro educativo en las poblaciones beneficiarias. Cabe agregar, que es sobre este pilar sobre el cual se fundamentan los procesos de evaluación dentro de los cuales se enmarca el desarrollo del presente TFM.

Laboratorios de Informática Educativa del PRONIE MEP-FOD

La estrategia de los LIE se fundamenta en las teorías del constructivismo y el construccionismo de Seymour Papert como base para desarrollar ambientes de aprendizaje que permitan a la población estudiantil involucrarse en actividades significativas, haciendo el aprendizaje real y susceptible de ser compartido (FOD, 2014b). En este espacio, se visualizan las tecnologías digitales “como una herramienta al servicio del desarrollo en las personas de capacidades cognitivas y sociales de alto nivel” (Muñoz et al., 2014, p.27).

Inicialmente los LIE funcionaron dentro de una propuesta pedagógica de Aprendizaje Basado en Proyectos (estrategia ABP) para orientar el trabajo en torno a la actividad de programación de computadoras. A partir del año 2009 se publicaron los *Estándares de desempeño de estudiantes en el aprendizaje con tecnologías digitales* (FOD, 2009) como un conjunto de perfiles que permitió precisar y articular de mejor manera los resultados de aprendizaje esperados en los y las participantes de las propuestas educativas del Programa. Bajo este modelo, se promueve que a través de herramientas tecnológicas digitales como las computadoras y algunas herramientas relacionadas con ambientes de programación, los y las estudiantes tengan la oportunidad de desarrollar competencias como la resolución de problemas, la investigación, la productividad, la ciudadanía y la comunicación (FOD, 2009; en adelante Estándares de desempeño).

Como se muestra en la **Figura 4**, dentro de los resultados de evaluación más recientes que se tienen de los LIE (FOD, 2017a), destaca que en las dimensiones de Ciudadanía y comunicación y la de Resolución de problemas e investigación es en las que se observan puntajes promedio mayores, mientras que en Productividad es en la que se obtienen puntajes menores. Según el estudio esto sugiere que en las primeras dimensiones los estudiantes tienen altas probabilidades de responder correctamente las tareas de nivel medio y bajo. Mientras que en el caso de la dimensión de Productividad, los estudiantes

tienen mayor probabilidad de responder correctamente a las tareas de dificultad baja (FOD, 2017a).

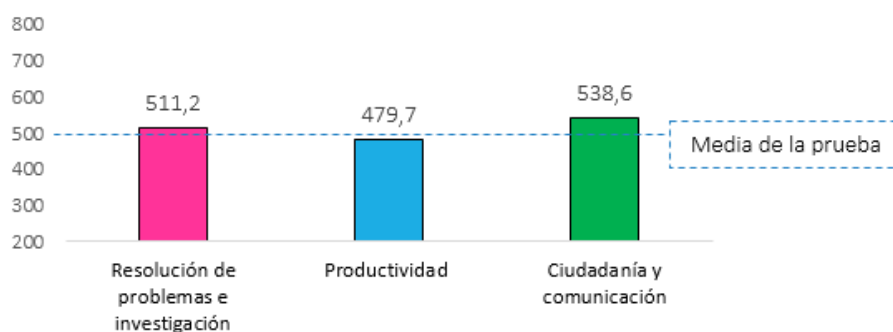


Figura 4. Nivel de logro de los estudiantes en las dimensiones de los Estándares de desempeño evaluados en el 2016 (n=20.548) (Fuente: FOD, 2017a, p. 31).

Otros hallazgos relevantes que se desprenden de este estudio, son (FOD, 2017a, pp 47-49):

- ✓ La principal oferta educativa que tienen los y las estudiantes evaluadas para aprender sobre TIC es la dada por el PRONIE MEP FOD. El aprendizaje que evidencian tener sobre tecnología y sobre las habilidades asociadas a las tres dimensiones se asocia positivamente con su participación en el espacio de los LIE.
- ✓ Se evidencia una tendencia al aumento de acceso a recursos tecnológicos como el celular e Internet, en comparación con los y las estudiantes evaluadas en años anteriores. Esto significa que la población estudiantil está cada vez más familiarizada con el uso de la tecnología en su hogar. Sin embargo, la tendencia de uso en ese contexto parece seguir orientándose al entretenimiento.
- ✓ Una mayoría de estudiantes muestra mejor comprensión del concepto de programación (en comparación con los estudios realizados en el año 2014 y el 2015), y más de la mitad reporta que podría programar sin ayuda. A su vez, los y las docentes tienen una valoración positiva de la programación como actividad de aprendizaje y consideran que brinda aportes y beneficios para la población estudiantil.
- ✓ Se identificó una asociación de variables como el sexo femenino, el no haber repetido ningún año en Primaria y el tener más años de participación en el Programa con mejores desempeños en las tres dimensiones estudiadas.

El siguiente paso en términos de los procesos de evaluación que se generan en este contexto de la Unidad de Evaluación de la FOD es poder clasificar a los y las estudiantes en niveles de logro o de competencia, en cada una de las dimensiones que componen el modelo de los Estándares de Desempeño. Dicho proceso es el que se pretende apoyar desde los objetivos planteados en el presente TFM.

II. Marco de referencia

A. Marco conceptual

Competencias digitales en el ámbito educativo

De acuerdo con Scolari (2018), desde que se inició el proceso de difusión de computadoras personales en la década de los 80 hasta su expansión mundial en los 90, así como el surgimiento de los teléfonos celulares y las redes sociales en los años 2000 “la tecnología digital ha sido la catalizadora del cambio social en las sociedades contemporáneas” (Scolari, 2018, p. 9). Evidentemente, este cambio ha permeado el campo educativo y ha modificado considerablemente las dinámicas e interacciones que se dan entre estudiantes y docentes.

Los cambios en las formas en las que se aprende y, en general, los cambios que las Tecnologías de la Información y la Comunicación (en adelante TIC) han supuesto para la educación han ido configurando nuevos perfiles de habilidades o competencias que se requieren, tanto por parte de la población estudiantil como del personal docente. En este contexto, surge el término de *competencia digital* en el ámbito educativo, entendido como:

(...) el uso crítico y seguro de las Tecnologías de la Sociedad de la Información para el trabajo, el tiempo libre y la comunicación; apoyándose en habilidades como el uso de ordenadores para recuperar, evaluar, almacenar, producir, presentar e intercambiar información, y para comunicar y participar en redes de colaboración a través de Internet (García-Varcárcel, 2016, pp. 3-4).

Como se observa en la **Figura 5**, esta definición supone que la competencia digital está compuesta por un conjunto de habilidades relacionadas con el manejo técnico de las tecnologías digitales, pero también de habilidades formales para utilizarlas y habilidades estratégicas, por ejemplo, para lograr un objetivo o una meta concreta (Pablos, Colás, Conde y Reyes, 2016). Según el modelo de estos autores, tres variables son clave dentro de la definición de la competencia digital, como son: 1) el uso personal y social de las tecnologías digitales; 2) la integración que se hace de éstas en las prácticas de los sujetos (internalización); y, 3) el bienestar emocional que el sujeto, en este caso estudiantes o docentes, tienen a partir de su uso.

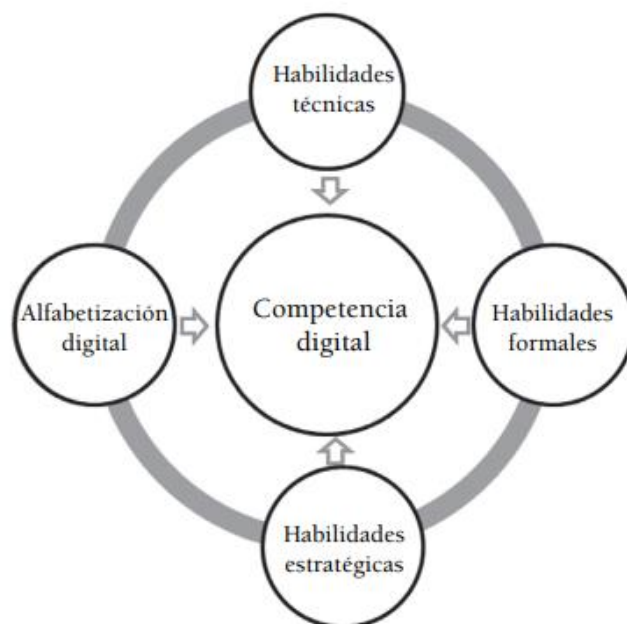


Figura 5. Modelo de dimensiones para la sistematización del concepto de competencia digital (Fuente: Pablos, Colás, Conde y Reyes, 2016, p. 6).

Pese a que la inclusión de tecnologías digitales en la educación es un movimiento que trasciende fronteras a nivel global, se han desarrollado diferentes modelos de conceptualización, según los objetivos y las prioridades que se han determinado en cada país (FOD, 2015b). Así, por ejemplo, algunos modelos como el de Singapur han desarrollado un concepto de competencia digital más orientado al planteamiento de metas de aprendizaje con el aprovechamiento de las tecnologías digitales dentro del currículo nacional, mientras que en otros, como el de Estados Unidos, se ha planteado como un marco complementario al currículo (Brenes et al., citados en FOD, 2015b).

A nivel global, se cuenta con diferentes estructuras para la conceptualización de la competencia digital. Sin embargo, también existen esfuerzos por unificar modelos y generar algunas pautas comunes. Al respecto, García-Varcárcel (2016) explica cómo desde el año 2011 en el contexto de la comunidad europea, se empieza a trabajar en la creación de un marco de referencia común sobre el desarrollo de la competencia digital. Para ello, dentro del marco del proyecto *European Digital Competence Framework for Citizen (DIGCOMP)* se definieron como objetivos:

1. Identificar los componentes claves de la competencia digital (conocimientos, habilidades y actitudes) necesarias para ser competente digitalmente.

2. Desarrollar los descriptores de la competencia digital para poder formular un marco teórico y poder validar diferentes niveles de competencia digital en Europa.
3. Proponer un plan de uso y desarrollo común de la competencia digital para diferentes niveles de aprendices.

Como resultado de este proyecto, se generaron algunos documentos clave que fueron útiles para conceptualizar el término de competencia digital, como fueron:

- ✓ **Mapping Digital Competence: Towards a Conceptual Understanding** (Ala-Mutka, 2011). Corresponde a una revisión teórica sobre el concepto de competencia digital y la pertinencia del desarrollo formal de ésta para la ciudadanía.
- ✓ **Digital Competence in practice: An analysis of Frameworks** (Ferrari, 2012). Presenta una amplia conceptualización correspondiente al estudio de 15 marcos específicos en los que se especifican diferentes niveles de alfabetización digital para Primaria, Secundaria, población adulta y la ciudadanía, en general.
- ✓ **Online Consultation on experts' views on Digital Competence** (Janssen y Stoyanov, 2012). Se basa en una consulta a expertos Europa a través de la técnica Delphi, lo que lleva a delimitar 12 áreas competenciales relacionadas con el uso de la tecnología para comunicarse, procesar y gestionar información y aprender, entre otros.
- ✓ **DIGCOMP: A Framework for Developing and Understanding Digital Competence in Europe** (Ferrari, 2013). Correspondiente a la publicación final del proyecto, se ofrece un marco conceptual útil para las iniciativas, currículos y certificaciones a nivel europeo sobre los componentes de la competencia digital. En este documento, define el término de la siguiente manera:

La competencia digital es un conjunto de conocimientos, habilidades, actitudes, estrategias, valores y concienciación (*dominios de aprendizaje*) que se requieren cuando se usan las TIC y los medios digitales (*herramientas*) para realizar tareas, solucionar problemas, comunicar, gestionar información, colaborar, crear y compartir contenido y construir conocimiento (*áreas competenciales*) de modo efectivo, eficiente, apropiado, crítico, creativo, autónomo, flexible, ético y reflexivo (*modos*) para el trabajo, el ocio, la participación, el aprendizaje, la socialización, el consumo y el empoderamiento (*propósitos*) (Ferrari, citado en García-Varcárcel, 2016, p.5)

Partiendo de este concepto, se genera un modelo que comprende un conjunto de habilidades y conocimientos organizados en cinco dimensiones y 21 competencias, las cuales se sintetizan en la **Figura 6** (García-Varcárcel, 2016).

1. Información:	<ul style="list-style-type: none"> • Identificar, localizar, recuperar, almacenar, organizar y analizar la información digital, evaluando su finalidad y relevancia.
2. Comunicación:	<ul style="list-style-type: none"> • Comunicar en entornos digitales, compartir recursos a través de herramientas en línea, conectar y colaborar con otros a través de herramientas digitales, interactuar y participar en comunidades y redes y desarrollar conciencia intercultural.
3. Creación de contenido:	<ul style="list-style-type: none"> • Crear y editar contenidos nuevos (textos, imágenes, videos...), integrar y reelaborar conocimientos y contenidos previos, realizar producciones artísticas, contenidos multimedia y programación informática, saber aplicar los derechos de propiedad intelectual y las licencias de uso.
4. Seguridad:	<ul style="list-style-type: none"> • Protección personal, protección de datos, protección de la identidad digital, uso de seguridad, uso seguro y sostenible.
5. Resolución de problemas:	<ul style="list-style-type: none"> • Identificar necesidades y recursos digitales, tomar decisiones a la hora de elegir la herramienta digital apropiada según la finalidad o necesidad, resolver problemas conceptuales a través de medios digitales, resolver problemas técnicos, uso creativo de la tecnología, actualizar la competencia propia y la de otros.

Figura 6. Dimensiones del Marco DIGCOMP para la conceptualización de la Competencia Digital a nivel europeo (Fuente: Ferrari, 2013).

En América Latina, las tecnologías digitales también se han incluido en el sistema educativo desde diferentes enfoques. Algunos han estado más orientados hacia el cierre de brecha de acceso como son los casos de Brasil, Perú, México y República Dominicana, mientras que otros como Chile y Costa Rica han estado más enfocados en desarrollar habilidades cognitivas y sociales en los estudiantes a través de su uso (FOD, 2015b). Según se explica:

En los casos de Chile y Costa Rica, con propuestas como *Enlaces* y la del *PRONIE MEP-FOD* (respectivamente), los objetivos de introducir las TIC en la educación han estado principalmente orientados a mejorar la calidad en el desarrollo de competencias que puedan ser aplicadas para resolver situaciones de la vida cotidiana. A partir de esta visión, en estos países la tecnología no ha sido el fin en sí

misma, sino que se ha concebido como herramienta para aprender y desarrollar habilidades como pensar, hacer y colaborar (FOD, 2015b, p.9).

Partiendo esa visión, dichos países han sido importantes referentes en la región en cuanto a la conceptualización de las habilidades o competencias que se desea desarrollar en los y las estudiantes (FOD, 2015b). En el caso de Chile, con el Marco SIMCE TIC y su respectiva *Matriz de Habilidades TIC para el aprendizaje de estudiantes* (Ministerio de Educación de Chile, 2013), se propone un marco útil para el diseño de políticas de inclusión y evaluación de la tecnología en la educación. Mientras que, en el caso de Costa Rica, el marco conceptual sobre el cual se lleva a cabo la evaluación de los y las beneficiarias de la propuesta de LIE del PRONIE MEP-FOD se define dentro del marco de los Estándares de Desempeño. Respondiendo al contexto dentro del cual se enmarca el presente TFM, en el siguiente apartado se detalla la conceptualización correspondiente al modelo costarricense.

Estándares de desempeño de estudiantes en el aprendizaje con tecnologías digitales del PRONIE MEP-FOD

Los Estándares de Desempeño “conjugan los diversos tipos de saber necesarios para la acción humana en contexto: el saber (*qué*), el saber hacer (*cómo*), el saber ser y el saber convivir (*por qué y para qué y para quiénes*)” (FOD, 2009, p. 10). Como se muestra en la **Figura 7** y en la **Tabla 1**, el modelo plantea una desagregación de tres dimensiones en siete estándares e integra de manera transversal una serie de propiedades como la creatividad y la colaboración, entre otros. Además, fue desarrollado por ciclo del sistema educativo, con lo cual se tiene un perfil de salida para Preescolar, *I y II ciclo* (correspondiente a los seis años de Primaria), y *III y IV ciclo* (correspondiente a los cinco años de Secundaria).

Como parte del interés institucional por alinear sus propuestas con las áreas de interés educativo a nivel internacional, durante el año 2015 se inició un proceso de análisis de concordancia entre las competencias digitales que se promueven desde los Estándares de Desempeño con otras propuestas educativas internacionales similares, como el marco de Habilidades del Siglo XXI del consorcio *Assessment and teaching of 21st century skills* (ATC21S, ver Griffin & Care, 2015). Según se describe en FOD (2017), este análisis permitió derivar de cada Estándar de desempeño una serie de actividades clave o hitos, que fueron útiles para construir la prueba de evaluación.

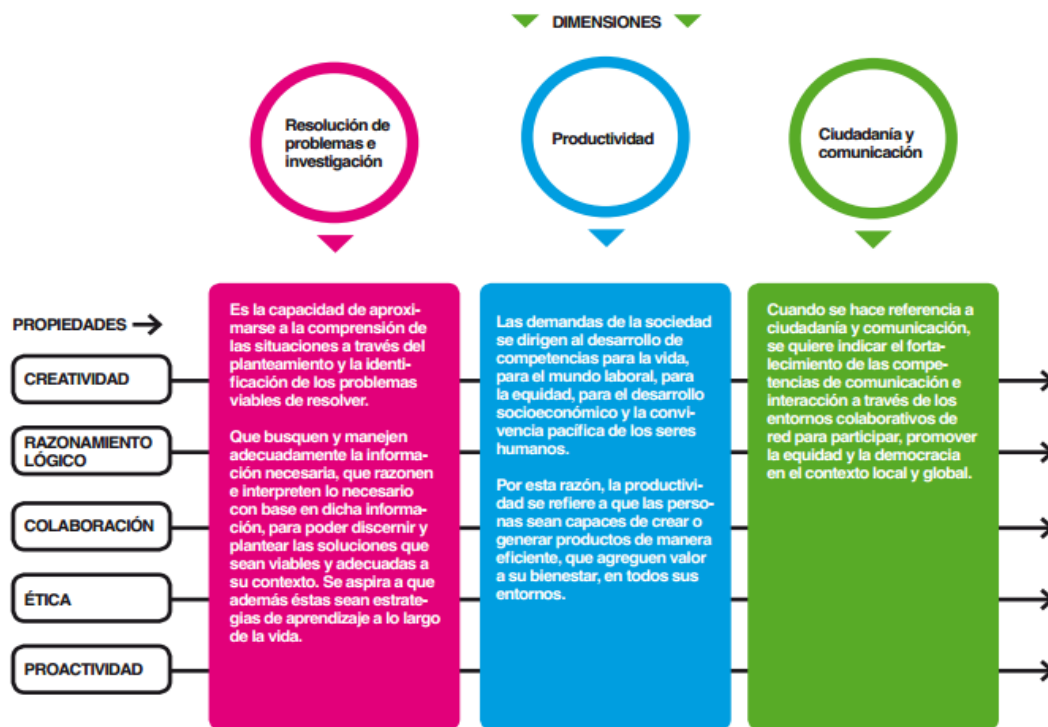


Figura 7. Dimensiones y propiedades de los Estándares de Desempeño del PRONIE MEP-FOD (Fuente: FOD, 2009, p. 12).

Tabla 1. Conceptualización del modelo de Estándares de Desempeño del PRONIE MEP-FOD.

Dimensión	Síntesis de los Estándares*
	<p>Estándar 1: Se enfoca en el desarrollo de producciones digitales.</p> <p>Estándar 7: Orientado al uso responsable de la tecnología.</p>
	<p>Estándar 2: Centrado en el desarrollo de un proyecto utilizando tecnologías digitales.</p> <p>Estándar 4: Orientado al desarrollo de productos programados.</p> <p>Estándar 5: Enfocado en la evaluación crítica de la información.</p>
	<p>Estándar 3: Centrado en la comprensión de entornos colaborativos.</p> <p>Estándar 6: Enfocado en la comprensión de las repercusiones de la tecnología en la vida de las personas.</p>

(Fuente: Elaboración propia a partir de la conceptualización de FOD, 2009).

Para el modelo de evaluación del año 2016 se derivaron 29 hitos de los siete estándares y, para cada hito, se construyó un conjunto de tres indicadores que posteriormente fueron traducidos en los ítems de la prueba (en total se construyeron 87 ítems). Como se muestra en la **Figura 8**, dicha desagregación permitió contar con una estructura conceptual útil para llevar a cabo la prueba de evaluación, sobre la que se fundamenta el interés de generar los puntos de corte adecuados para ofrecer resultados sobre el nivel de habilidad de los y las estudiantes.

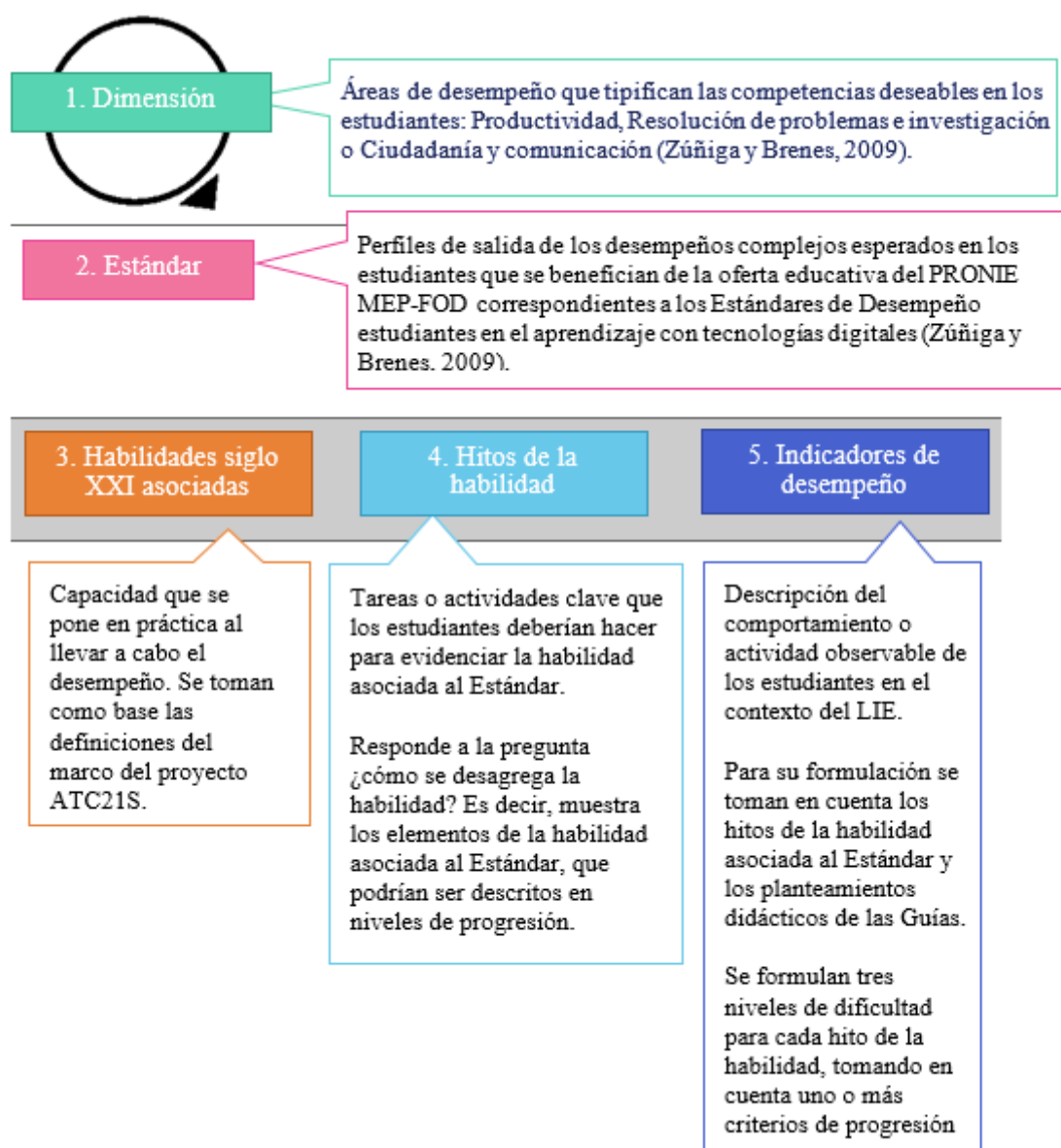


Figura 8. Detalle de la estructura conceptual para la evaluación de los Estándares de Desempeño del PRONIE MEP-FOD (Fuente: Villalobos, Núñez, Sequeira y Brenes, 2018, p. 9).

Según se detalla en el diseño del estudio vigente para el año 2018 (FOD, 2018), durante este año se priorizará la evaluación de dos de las tres dimensiones de los Estándares de Desempeño, las cuales corresponden a Productividad y Resolución de problemas e Investigación. Partiendo de esto, el marco de indicadores se reduce a 63 que corresponde a 21 de los 29 hitos originalmente derivados. El detalle de dicha desagregación de indicadores a evaluar se muestra en la **Figura 9**.

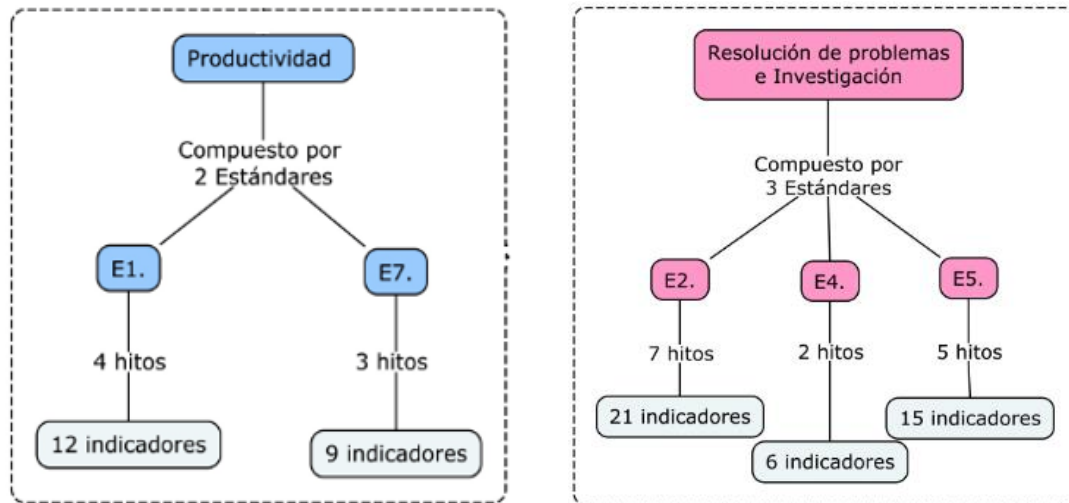


Figura 9. Distribución de los Estándares de Desempeño e hitos en las dimensiones a evaluar en el 2018 [*E*= estándar]. (Fuente: FOD, 2018, p. 24).

Sobre la prueba de evaluación es importante mencionar que sus ítems se caracterizaron por ser de selección múltiple con respuesta única, con un formato de enunciado con cuatro opciones de respuesta. Estos fueron creados con base a tres niveles de progresión, correspondientes a los indicadores de evaluación. Según se detalla en FOD (2018), se espera que la prueba tenga una duración aproximada de 60 minutos y que esté disponible tanto de forma digital (en la plataforma en línea *LimeSurvey*) como en formato impreso, para garantizar la aplicación aún en contextos con condiciones limitadas de conectividad.

Finalmente, es importante mencionar que el objetivo de evaluar el nivel de logro de los y las estudiantes beneficiarias del PRONIE MEP-FOD responde a un interés por dar cuenta de sus resultados de aprendizajes y estimar el aporte del Programa a los mismos. Por ello, se requiere de una estrategia de definición de puntos de corte que permita identificar varios niveles de competencia. Ante esto, se considera que aquellas estrategias cuyo resultado es un punto de corte binario (tipo “aprobado” o “no aprobado”) no serían relevantes para este contexto de evaluación.

Evaluación referida a criterios y establecimiento de puntos de corte

Además del modelo conceptual sobre el cual se fundamenta la prueba de evaluación de los y las estudiantes beneficiarias de los LIE, se considera oportuno hacer referencia a las pruebas de evaluación referidas a criterios y a la conceptualización general de las estrategias de definición de puntos de corte. Se toma como punto de partida la diferenciación entre las pruebas referidas a criterios con las que se construyen más tradicionalmente, que son las pruebas referidas a la norma. Desde principios de la década de los 80, Popham (1983) hacía referencia a la diferencia al definir las de la siguiente manera:

Un test basado en pautas de normalidad está destinado a determinar la posición de un sujeto examinado en relación con el rendimiento de un grupo de otros sujetos que hayan hecho ese mismo test (Popham, 1983, p. 47).

Un test basado en criterios se emplea para determinar la posición de un individuo con respecto a un dominio de la conducta perfectamente definido (Popham, 1983, p.134).

Partiendo de estas definiciones, se observa que la principal diferencia entre ambos tipos de prueba radica principalmente en el referente con el cual se contrastan los resultados de los participantes. En el primer caso (con las pruebas referidas a la norma), se busca establecer comparaciones de los resultados de cada participante con respecto a su propio grupo. Mientras que, en el segundo caso (con las pruebas referidas al criterio), se busca identificar el nivel o posición que muestra cada participante en relación con determinado nivel de competencia o dominio.

La diferencia descrita se considera base para la identificación de distinciones más específicas como, por ejemplo, en cuanto al objetivo que se busca con cada tipo de prueba. De acuerdo con Barrios y Coscolluela (2013), en el caso de las pruebas referidas a la norma el principal objetivo es la medición de una variable psicológica o un rasgo, con lo cual tienden a ser más empleadas en ámbitos de evaluación de personalidad o de actitudes. Mientras que las pruebas referidas al criterio, tienen como principal objetivo medir un conjunto de conocimiento o competencias, por lo que tienden a ser más empleadas en el ámbito de la educación y de la certificación de competencias.

Además de la diferencia vinculada al objetivo se plantean otros aspectos de los cuales es posible observar distinciones entre ambos tipos de pruebas. Al respecto, en la **Tabla 2** se

muestra una síntesis de los principales aspectos destacados por autores como Esquivel (2001) y Barrios y Cosculluela (2013):

Tabla 2. Diferencias entre las pruebas referidas a normas y las pruebas referidas a criterios

	Pruebas referidas a normas	Pruebas referidas a criterios
Paradigma de base	Paradigma psicométrico. Busca normalidad en los resultados	Paradigma educométrico. Busca que la mayoría de sujetos aprenda
Objeto que evalúa	Miden una variable psicológica o un rasgo	Miden un conjunto de conocimientos o competencias
Análisis de fiabilidad	A partir de las diferencias entre sujetos	A partir del dominio que presenta el individuo sobre el tema
Objetivo de la evaluación	Posición relativa de un individuo con respecto al resto del grupo	Grado de conocimiento que presentan los individuos en el dominio o competencia
Ámbito de aplicación	Principalmente en pruebas de personalidad, actitudes, etc.	Principalmente en pruebas de Educación, ámbito laboral, evaluación de programas, etc.
Interpretación de resultados	Es relativa. El puntaje tiene sentido al ser comparado con medidas de tendencia central del grupo. No requiere establecer puntos de corte sino que el resultado se interpreta en función de los resultados del grupo.	Es absoluta. Se interpreta en términos de logro o no logro, o de ubicación en determinado nivel de competencia. Requiere establecer puntos de corte para clasificar a los sujetos según niveles de logro o de competencia, de manera independiente al grupo.

(Fuente: Barrios y Cosculluela, 2013, p.113; Esquivel, 2001, pp. 26-27).

En materia específica de pruebas de evaluación referidas a criterios, es importante hacer referencia a cuatro operaciones fundamentales que se deben tomar en cuenta para su construcción (Aranguren y Hoszowski, 2017):

1. La especificación del dominio o contenido que se quiere evaluar.
2. El diseño y análisis de los ítems que se incluirán en la prueba.
3. El establecimiento de los puntos de corte y la definición de los niveles de logro.
4. La documentación y análisis de las evidencias de confiabilidad y validez de las pruebas.

Partiendo de lo anterior y en concordancia con el foco del presente TFM, se hará referencia a la noción de establecimiento de los puntos de corte, es decir, a cómo se lleva a cabo la definición de los niveles mediante los cuales se pretende clasificar o ubicar el logro de los participantes en la prueba (Cizek & Bunch, 2007). Una definición adecuada de lo que es establecer un punto de corte la brindan estos autores, de la siguiente manera:

(...) se puede definir como un procedimiento que permite a los participantes que utilizan un método específico llevar a cabo sus juicios de forma tal que puedan traducir las posiciones políticas de las entidades autorizadas a ubicaciones en una escala de puntaje [traducción libre] (Cizek & Bunch, 2007, p.19).

Los aspectos psicométricos que involucran las estrategias de definición de puntos de corte son útiles para asegurar que las decisiones sobre los resultados de un grupo de evaluados se lleven a cabo de manera sistemática, reproducible y objetiva (Cizek & Bunch, 2007). Berk (citado en Barrios y Cosculluela, 2013), concuerda con esta idea, sin embargo, considera que el proceso inevitablemente involucra un balance entre las perspectivas basadas en el establecimiento de estándares de logro, las agendas políticas, las presiones económicas y las prioridades a nivel de política educativa, entre otros factores. Es por esto que se considera que la definición de un punto de corte que sea satisfactoria para todos los actores involucrados difícilmente existe (Hofstee, citado en Cizek & Bunch, 2007).

Desde hace varias décadas se han desarrollado numerosas estrategias que pretenden responder a las demandas de definición de puntos de corte en diferentes contextos, con lo cual actualmente existen diversas opciones para generarlos. Es por ello que se han generado diferentes clasificaciones de las estrategias disponibles, dentro de las que destaca la de Barrios y Cosculluela (2013, p.127) y respaldada por Muñiz (2018):

1. **Estrategias centradas en el test (*Test Centered*)**. Corresponde a una serie de métodos o estrategias basadas en la valoración que un grupo de expertos hace de manera detallada sobre los ítems de una prueba de evaluación. Dentro de este grupo destacan estrategias como las de Nedelsky (1954), Angoff (1971) y las de Sireci, Hambleton y Pitoniak (2004).
2. **Estrategias centradas en los examinados (*Examinee Centered*)**. Corresponde a métodos basados en la valoración que un grupo de expertos hace sobre la competencia o desempeño de los sujetos evaluados. Destacan estrategias como las de Grupos de

contraste de Berk (1976) y las de Grupos límite de Zieky y Livingston (1977). Cabe mencionar que Cizek y Bunch (2007) también denominan este tipo de estrategias como *métodos holísticos*².

- 3. Métodos de compromiso (*Norm referenced*).** Dentro de los que destacan las estrategias que equilibran la emisión de los juicios con la comparación del sujeto con su grupo normativo, es decir, con datos empíricos. Se incluyen dentro de este grupo técnicas como las de Hofstee (1983) y Beuk (1984) y otras desarrolladas más recientemente como es el caso del método *Bookmark* (Jornet y Gonzales, 2009).

Como característica común, una de las principales ventajas que tienen las estrategias de definición de puntos de corte en pruebas referidas a criterios es que se determinan los niveles de habilidad de los evaluados de manera más cercana a la realidad al tomar en cuenta el criterio de expertos en el contenido (McClarty, Way, Porter, Beimers, & Miles, 2013). Dicha característica cobra relevancia en el presente TFM, al considerar que en este caso, se trata de la definición de puntos de corte para una prueba enmarcada en el ámbito de la educación en Costa Rica en el que interesa implementar una estrategia que sea más cercana a la realidad educativa.

Otra característica común es que las estrategias para definir puntos de corte en pruebas de evaluación referidas a criterios utilizan diferentes tipos de información como estrategia para dar realimentación a los jueces expertos durante las jornadas de trabajo. Aunque cada estrategia emplea diferente información según sus requerimientos, en la **Tabla 3** se muestran los tres principales tipos de información que son posibles de utilizar en el proceso, según Cizek y Bunch, 2007).

² Implican una valoración categórica de los evaluados y sus resultados reales en la prueba, con lo cual el punto de corte depende del desempeño de cada muestra analizada (Cizek y Bunch, 2007).

Tabla 3. Tipos de información utilizada para dar realimentación a los jueces

Información normativa	Información de realidad	Información de impacto
<ul style="list-style-type: none"> • Puntos de corte por participante (por juez) • Distribución de los puntos de corte emitidos por todos los jueces. • Matriz de puntuación de cada participante. 	<ul style="list-style-type: none"> • P values o b values de los ítems (o cualquier otra escala de valores) • Probabilidad de respuesta (por ejemplo RP50 o RP67) • Tablas de distribución de los resultados • Calificaciones de los participantes con la variable de criterio externo (por ejemplo, juicios de director de programa, calificaciones de supervisor u otros datos) 	<ul style="list-style-type: none"> • Distribuciones de puntajes en bruto (o escalados) • Porcentajes de examinandos en cada categoría de rendimiento

(Fuente: Elaboración propia a partir de Cizek y Bunch, 2007).

Una de las principales interrogantes que han surgido dentro de este campo de estudio ha sido el cómo identificar cuál estrategia podría ser más pertinente. Al respecto, García, Abad, Olea y Aguado (2012) explican que, de acuerdo con la Asociación Americana de Investigación Educativa (*American Educational Research Association –AERA*, 1999) no es posible definir un único procedimiento para definir los puntos de corte, ya que un mismo método puede ser aplicado de diferentes maneras según los requerimientos contextuales.

En esta misma línea, Cizek y Bunch (2007) consideran que la estrategia por sí misma no define un buen o mal procedimiento de establecimiento de puntos de corte, sino que esto depende más del conjunto de decisiones que se adoptan alrededor de su implementación. Aun cuando no se defina un método cómo mejor otro, existen algunos criterios útiles para orientar su elección. Al respecto, Norcini (2003) hace referencia a pasos generales que se pueden considerar, como es la necesidad de decidir el tipo de estándar que se quiere evaluar y la estrategia, antes de reclutar a los jueces. Posteriormente, recomienda llevar a cabo los procesos de entrenamiento presencial de éstos para calcular los puntos de corte y, por último, establecer qué acciones concretas realizar en función de los resultados obtenidos.

En la **Figura 10** se muestra una síntesis de los aspectos genéricos que Cizek y Bunch (2007) recomiendan para seleccionar una estrategia de definición de puntos de corte:

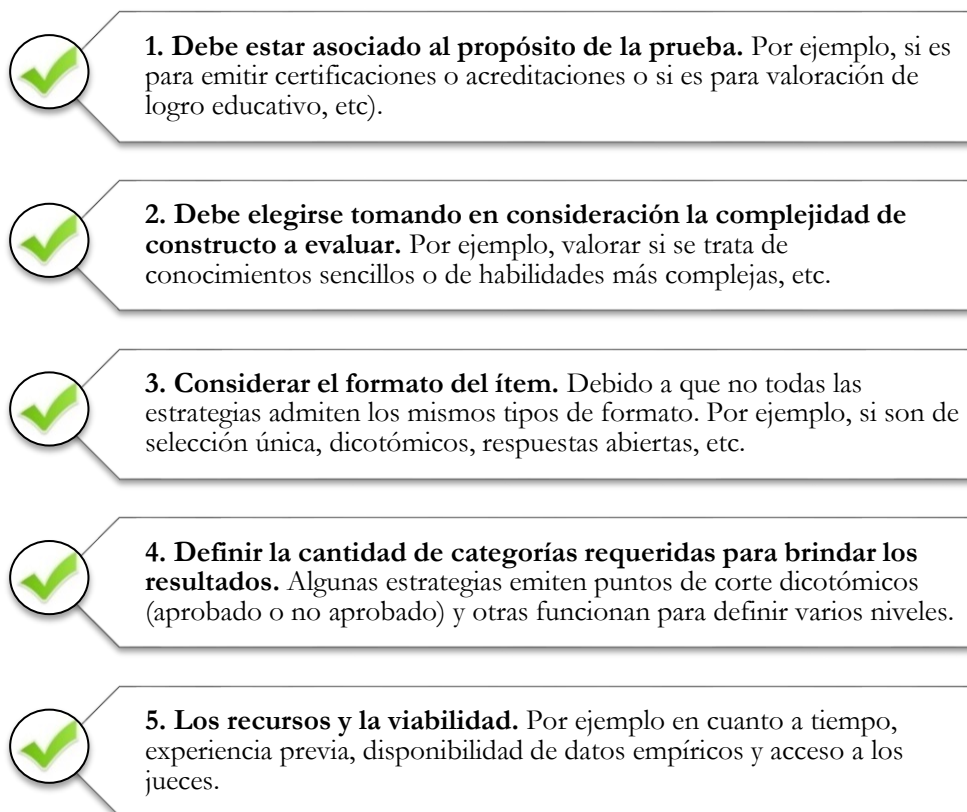
- 
- 1. Debe estar asociado al propósito de la prueba.** Por ejemplo, si es para emitir certificaciones o acreditaciones o si es para valoración de logro educativo, etc).
 - 2. Debe elegirse tomando en consideración la complejidad de constructo a evaluar.** Por ejemplo, valorar si se trata de conocimientos sencillos o de habilidades más complejas, etc.
 - 3. Considerar el formato del ítem.** Debido a que no todas las estrategias admiten los mismos tipos de formato. Por ejemplo, si son de selección única, dicotómicos, respuestas abiertas, etc.
 - 4. Definir la cantidad de categorías requeridas para brindar los resultados.** Algunas estrategias emiten puntos de corte dicotómicos (aprobado o no aprobado) y otras funcionan para definir varios niveles.
 - 5. Los recursos y la viabilidad.** Por ejemplo en cuanto a tiempo, experiencia previa, disponibilidad de datos empíricos y acceso a los jueces.

Figura 10. Síntesis de aspectos a considerar para elegir una estrategia de definición de puntos de corte. (Fuente: Elaboración propia a partir de las recomendaciones de Cizek y Bunch, 2007).

Algunos aspectos que los autores destacan como claves para incrementar la confiabilidad en un proceso de definición de puntos de corte son asegurarse de que la estrategia elegida efectivamente se relaciona con las demandas y características de la prueba, así como el identificar y seleccionar un adecuado conjunto de jueces y brindarles un adecuado entrenamiento para optimizar los juicios que emiten (Cizek y Bunch, 2007). Muñiz (2018) menciona que la cantidad de jueces debe ser representativa y cuanto más amplia mejor, pero no se define un número concreto para ninguna estrategia.

Se evidencia que la estrategia de definición de puntos de corte que se elija para un determinado contexto va a estar directamente relacionada con las características propias del modelo de evaluación al que responde. Con ello, se sostiene el objetivo del presente TFM orientado a comparar las estrategias disponibles en función de la pertinencia para la evaluación de competencias digitales de estudiantes en el ámbito educativo costarricense.

Science Mapping como estrategia de revisión creativa de la literatura

Hart (1998) hace referencia a la noción de revisión creativa de la literatura al mencionar la necesidad de emplear habilidades cognitivas de orden superior para elaborar búsquedas de información que sean oportunas, competentes y, principalmente, que faciliten la exploración de un campo de conocimiento. Para ello, define el concepto alrededor de la necesidad de profundizar para tener una visión amplia de un tema a partir de la posibilidad de explorar diferentes ideas para buscar relaciones no evidentes entre los datos disponibles.

Con el objetivo de explorar un campo de conocimiento desde esta perspectiva de revisión creativa de la literatura, Hart (1998, p. 30) define una serie de interrogantes clave a considerar como punto de partida, las cuales corresponden a:

1. ¿Cuál es la estructura del conocimiento en este tema?
2. ¿Cuáles son las palabras clave de los principales trabajos y quiénes son los teóricos que lo respaldan?
3. ¿Cuáles son las premisas metodológicas y morales sobre las cuales se fundamenta el estudio de este campo de conocimiento?
4. ¿Cómo se relacionan entre sí los diferentes estudios identificados en el tema?

Las interrogantes muestran una aproximación inicial a partir de la cual se puede llevar a cabo un proceso de revisión creativa de la literatura sobre un tema de interés. Para llevarlo a cabo, el autor define una serie de estrategias oportunas a partir de las cuales se podría abordar un proceso de revisión bibliográfica desde este enfoque. Al respecto, destacan estrategias de argumentación y análisis, técnicas de organización y expresión de ideas y estrategias de mapeo o *Science Mapping* como estrategias que permiten explorar una idea, una técnica o una teoría desde un enfoque de revisión creativa de la literatura (Hart, 1998).

Rodríguez-Bolívar, Alcaide-Muñoz y Cobo (2018), indican que la estrategia de *Science Mapping* es útil para explorar un campo de conocimiento desde una perspectiva estructural y dinámica, ya que permite explorar su evolución en el tiempo, mapear la literatura disponible y trabajar directamente con conjuntos de términos compartidos entre los documentos publicados, entre otras ventajas. Asimismo, la estrategia ofrece la oportunidad de “mapear las relaciones intelectuales que existen en la literatura sobre un tema determinado y así revelar los patrones implícitos, entre los orígenes de una idea, su

desarrollo y su implementación” [traducción libre] (Hart, 1998, p. 161). De esta manera, se justifica la elección de esta estrategia metodológica como hilo para el desarrollo del TFM.

Para la implementación de una estrategia de revisión creativa de la literatura desde el enfoque de *Science Mapping* se propone un procedimiento que involucra un flujo de trabajo desde el momento en el que se recopilan los datos hasta que se interpretan. Al respecto, Rodríguez-Bolívar, Alcaide-Muñoz & Cobo (2018) proponen el esquema que se observa en la **Figura 11**, dentro del cual se contemplan aspectos vinculados con la recolección y depuración de la fuente de datos, así como con el análisis e interpretación de los resultados.

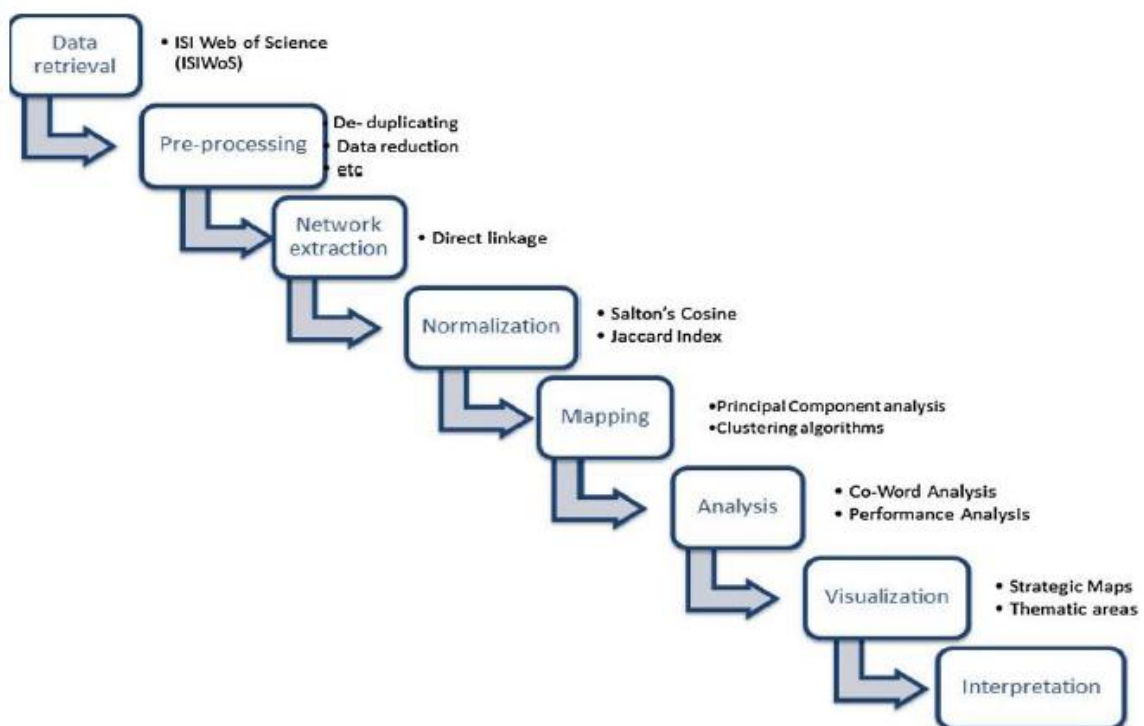


Figura 11. Diagrama de flujo para la implementación de la estrategia de *Science Mapping* (Fuente: Rodríguez-Bolívar, Alcaide-Muñoz & Cobo, 2018, p. 113).

Este tipo de procedimientos son posibles gracias a las ventajas que ofrecen actualmente las tecnologías digitales y la web 2.0. Para ello, se han generado diferentes herramientas de apoyo que facilitan la navegación y exploración de los patrones y las tendencias que se dan a lo interno de un conjunto de publicaciones científicas (Chen, 2017). Cada herramienta tiene sus respectivas funcionalidades y limitaciones según el objetivo investigativo con el que se aborda un campo de conocimiento. Algunas de las que están disponibles en la actualidad son *VOSViewer* (véase el detalle en Van Eck & Waltman, 2018), *CiteSpace* (véase el detalle en Chen, 2018) y otras como *HistCite*, *SciMAT*, y *Sci2*.

B. Antecedentes del estudio

La definición de puntos de corte en pruebas referidas a criterios es un campo de conocimiento que ha sido ampliamente estudiado desde hace varias décadas (Cizek y Bunch, 2007). Por una parte, se cuenta con estudios que se han enfocado en valorar un solo método de definición de puntos de corte, con el fin de aportar validez empírica y de probar el funcionamiento de algunas variaciones propuestas. Mientras que, por otra parte, se cuenta con amplia bibliografía alrededor de la comparación de resultados entre las diferentes estrategias.

Dentro de los hallazgos más recientes en torno a la exploración de un solo método destaca la investigación de Margolis y Clauser (2014) en la que se evaluó el impacto de una modificación al método *Angoff* original, correspondiente a la provisión de datos sobre el rendimiento del examinado, es decir, de datos empíricos. Para ello, se examinaron los datos de tres paneles de jueces (con un promedio de 10 jueces por panel) en tres exámenes de licencias médicas para investigar si la provisión de dicha información impactaba los juicios emitidos. Con ese propósito, se asignó a cada juez un conjunto de ejercicios representativos de una prueba para certificación de competencias en el campo de la medicina (*The United States Medical Licensing Examination –USMLE*).

Cada juez revisó los ítems y emitió su primer juicio según las indicaciones correspondientes al método *Angoff* original. Posteriormente, recibieron información de realidad, es decir, acerca del desempeño de los y las evaluadas, así como los porcentajes de selección de cada opción de respuesta y tuvieron la oportunidad de ajustar sus juicios. Aunque los resultados variaron según el panel, en general, tanto la variabilidad entre los jueces como los puntajes de corte resultantes del proceso se vieron afectados por los datos empíricos provistos (Margolis & Clauser, 2014). Por ejemplo, después de la revisión de los datos de rendimiento en la prueba, la variabilidad de los jueces disminuyó de manera significativa.

De acuerdo con los autores, el estudio apoya con evidencia empírica la premisa de que la interacción entre los jueces y la información que provee cada método tiene una influencia en la variabilidad de los puntos de corte resultantes. Estudios similares se han planteado con métodos concretos como el *Bookmark*, por ejemplo en el establecimiento de puntos de corte para pruebas de idiomas (Tiffin-Richards & Anand, 2013), cuyos principales hallazgos muestran que la selección de los ítems que se utilizan para el trabajo con el cuadernillo, así

como la revisión que los jueces hacen de sus primeros juicios emitidos parecen tener efectos sobre la variabilidad del establecimiento final de los puntos de corte.

Aun cuando se desarrollan este tipo de investigaciones centradas en un solo método, la corriente investigativa que se ha seguido de manera más fuerte en los últimos años corresponde a la comparación de los resultados obtenidos a través de la aplicación de diferentes métodos. Bajo la premisa de necesitar mayor investigación en el campo, algunos de los principales estudios se orientaban a respaldar procedimientos como el *Angoff* y el de *Grupos límite* como enfoques más razonables y defendibles en comparación con métodos más relativos y holísticos (Kaufman, Mann, Muijtjens & Van der Vleuten, 2000).

Buckendahl, Smith, Impara y Plake (2002), llevaron a cabo una comparación entre dos de los métodos más utilizados en la actualidad, los cuales corresponden precisamente al método *Angoff* (como uno de los más antiguos y estudiados) y al método *Bookmark* (como uno de los más recientemente desarrollados). Para el establecimiento de la comparación se tomaron los resultados de una prueba de evaluación aplicada a una muestra de 448 estudiantes de séptimo grado, desarrollada para evaluar habilidades en matemáticas. Específicamente, la prueba incluía ítems de selección única que aproximaban seis áreas de habilidades matemáticas como álgebra, geometría y análisis de datos, entre otras (Buckendahl et al., 2002).

En primer lugar, luego del correspondiente entrenamiento de los jueces se les solicitó revisar cada uno de los ítems de la prueba y determinar si un grupo de sujetos con un mínimo nivel de competencia sería capaz de contestarlo de manera correcta (aplicación del método *Angoff*). En segundo lugar, se les brindó un cuadernillo de trabajo en el que encontraron los ítems ordenados según el nivel de dificultad que tuvieron para los estudiantes. Bajo la consigna de pensar en un grupo de estudiantes menor habilidad, los jueces colocaron marcas en los ítems en los que consideraban que los sujetos dejarían de acertar las respuestas según una norma de probabilidad de acierto establecida previamente (aplicación del método *Bookmark*).

Se encontraron variaciones entre los puntos de corte definidos por cada método (véase el detalle de las variaciones en la **Figura 12** y la **Figura 13**). Según explican los autores, en el caso del método *Angoff*, las variaciones entre cada ronda de revisión por parte de los jueces podría deberse a que inicialmente se apartaron de su conceptualización del grupo de estudiantes con mínima habilidad y se enfocaron más en sus creencias personales sobre el

puntaje de corte apropiado (Buckendahl et al., 2002). Mientras que, en el caso del método *Bookmark*, se hipotetiza que cuando los jueces tienen información sobre el impacto que sus puntos de corte podrían tener en la población evaluada, tienden a reducir la varianza en los criterios emitidos, lo cual se considera uno de los mejores indicadores para seleccionar una estrategia.

TABLE 1
Angoff method cut score means¹ and standard deviations for Rounds 1 and 2

Round	Cut score	SD	Percent failure
1	34.92	7.79	8.9
2	33.42	10.96	7.6

¹ Based on 69 total points

Figura 12. Tabla de resultados de los puntos de corte establecidos mediante el método *Angoff* (Fuente: Buckendahl et al., 2002, p. 259).

TABLE 2
Bookmark method cut score means¹ and standard deviations for Rounds 1 and 2

Round	Cut score	SD	Percentage failure
1	33.64	11.03	7.6
2	35.64	8.66	9.4

¹ Based on 69 total points

Figura 13. Tabla de resultados de los puntos de corte establecidos mediante el método *Bookmark* (Fuente: Buckendahl et al., 2002, p. 260).

Una de las principales conclusiones a las que llegan los autores con este estudio comparativo, es que el método *Bookmark* constituye una estrategia prometedora para el campo de la educación en tanto que su metodología reduce muchas de las preocupaciones que se han mantenido a lo largo de los años alrededor de otros métodos tradicionales como el *Angoff*. Por ello, enfatizan en la necesidad de continuar explorando y documentando experiencias de implementación de esta estrategia (Buckendahl et al., 2002).

La línea de comparación entre estos dos métodos se ha mantenido, con lo cual se ha documentado de manera importante en los últimos años (Cetin & Gelbal, 2013; Hamme & Shulz, 2011; Reckase, 2006; Sanju, Sayeed & Femi, 2006). Al respecto, destaca la llevada a cabo por Hamme y Shulz (2011), en la que se realizó una revisión de las publicaciones hechas en los últimos 15 años, en las que se comparaban ambas estrategias.

Según explican los autores, históricamente en Estados Unidos el Sistema Nacional de Evaluación del Progreso Educativo (*National Assessment of Educational Progress* –NAEP) ha trabajado con estrategias de definición de puntos de corte basadas en los métodos *Angoff*. Sin embargo, a partir del año 2005 se valoró la información que algunos estudios comparativos empezaban a arrojar en términos de la confiabilidad, la validez y las características de los procedimientos de dicho método en comparación con el método *Bookmark*. Considerando que éste último se colocaba como un método con características más favorables, se adoptó como estrategia de definición de puntos de corte en el NAEP (Hamme & Shulz, 2011).

Los autores fundamentan su investigación en un análisis en profundidad de 27 artículos científicos que fueron base para la decisión del NAEP de cambiar su método. De esta manera, centraron su atención en valorar los hallazgos correspondientes a ciertos aspectos clave como la validez del punto de corte, la variación entre un método y otro y la confiabilidad del procedimiento de trabajo con los jueces.

Entre las principales conclusiones destaca que el método *Bookmark* tiene mejor evidencia de fiabilidad y validez. Se identificó que las desviaciones estándar obtenidas de las puntuaciones asignadas por los jueces mediante este método, son menores a las que se obtienen del método *Angoff*, lo que sugiere una mayor confiabilidad. Además, los puntos de corte resultantes de ambos métodos convergieron para la misma área de contenido, evidenciando mayor validez en el caso del método *Bookmark*. Finalmente, la comprensión de los jueces sobre las tareas y las instrucciones fue más alta en el caso de este último método (Hamme & Shulz, 2011).

Dentro de los ejemplos de otros estudios comparativos, Van Niljen y Janssen (2008) llevaron a cabo una comparación empírica entre el método *Angoff* como representante del grupo de estrategias de definición de puntos de corte centradas en el ítem y el método de *Grupos de contraste* como representante del grupo de estrategias centradas en el examinado. Para ello, utilizaron los resultados de una prueba que evaluaba competencias en biología de la población de Primaria de la región de Flandes en Bélgica. De acuerdo con los autores, se trabajó con los resultados de 1,321 estudiantes (pertenecientes a 61 centros educativos) y la prueba estaba compuesta por un total de 162 ítems de formato mixto (de selección única y de respuesta abierta). Para la comparación de las estrategias de definición de puntos de

corte, se contó con participación de 69 docentes para el método *Angoff* y de 73 para el caso del método de *Grupos de contraste* (Van Niljen & Janssen, 2008).

La primera tarea a la que se enfrentaron los jueces fue a clasificar a sus estudiantes como *masters* y *nonmasters*, es decir, que tuvieron que dividir su respectivo grupo de estudiantes en aquellos que consideraban que tenían un dominio suficiente del contenido y los que no (aplicación del método de *Grupos de contraste*). La segunda tarea, consistió en revisar cada uno de los ítems de la prueba y determinar si un sujeto con el mínimo nivel de competencia en biología sería capaz de responderlo o no (aplicación del método *Angoff*).

Dentro de los principales resultados destaca una baja correlación entre la identificación de los puntos de corte obtenidos con ambos métodos, siendo el método *Angoff* con el que se obtuvieron puntos de corte más altos (Van Niljen & Janssen, 2008). Una de las posibles explicaciones que dan los autores a este hallazgo, es que la tarea para el juez en esta estrategia involucra valorar cuál ítem *debería* (*should*) ser respondido por un estudiante con mínimo nivel de competencia y esta es una palabra que suele interpretarse como un máximo objetivo deseado.

Otro hallazgo importante es que los resultados de la regresión evidenciaron que los juicios emitidos estaban asociados de manera significativa con la capacidad de los estudiantes en el dominio de la biología. Este hallazgo resta credibilidad a una de las principales críticas que se hace al método de *Grupos de contraste*, relacionado con el sesgo de criterio que podrían tener los jueces al conocer directamente a los y las evaluadas (Van Niljen & Janssen, 2008).

Los autores concluyen que el método que se seleccione para definir puntos de corte en una prueba tiene un efecto en la determinación éstos. Dicha conclusión sugiere la existencia de una interacción entre el juez y el método seleccionado, lo cual apoya la idea de que es difícil llegar a un acuerdo entre diferentes contextos sobre cuál estrategia es más pertinente (Jaeger, 1989; Van Niljen & Janssen, 2008; Zieky, 2001).

Finalmente, en la misma línea de estudios comparativos cabe hacer referencia al trabajo de Aranguren y Hoszowski (2017), quienes tomaron varios estudios en los que se habían valorado algunas estrategias de establecimiento de puntos de corte según un modelo de criterios de valoración planteado por Berk (1986). Los autores comparan los hallazgos de los diferentes artículos y elaboran la síntesis que se muestra en la **Figura 14**, la cual

evidencia que el método *Bookmark* es el que obtuvo mejores puntuaciones en la mayoría de los componentes estudiados³.

Criterios de Berk (1986)		Angoff (Ricker, 2006)	Bookmark (Lin, 2006)	BoW(Radwan & Rogers, 2006)	Juicio analítico (Abbott, 2006)
De adecuación técnica	Clasificación apropiada de los datos	2	3	3	3
	Sensible al desempeño de los examinados	2	3	3	2
	Consideración de aspectos relacionados con la instrucción y entrenamiento de los jueces	1	3	3	3
	Aporte de datos y sustento estadístico para la toma de decisiones	3	3	2	3
	Identificación del "verdadero" estándar	2	3	3	1
	Aporte de evidencias de validez para las decisiones	2	2	1	1
Factibilidad	Fácil de implementar	2	3	2	2
	Fácil de computar	3	3	3	3
	Fácil de interpretar	3	3	2	3
	Confiable para los no especialistas en el área	3	3	2	2

Figura 14. Comparación entre las metodologías *Angoff*, *BoW* (*Cuerpo de Trabajo*), *Juicio analítico* y *Bookmark* siguiendo criterios de Berk (1986). (Fuente: Aranguren y Hoszowski, 2017, p. 33)

Al hacer una valoración global de los principales antecedentes con los que cuenta el campo de conocimiento, es posible identificar que, no es posible identificar una estrategia de definición de puntos de corte como la más adecuado, sino que la elección debe hacerse en función de los objetivos de la evaluación y de las características metodológicas y contextuales a partir de las cuales se diseña el modelo evaluativo. Lo que sí parece estar claro es que, cualquier controversia en relación con la selección de una u otra estrategia podría minimizarse implementándola fielmente y reuniendo evidencia sólida sobre la validez del proceso y de los resultados obtenidos (Cizek, Bunch y Koons, 2004).

³ La escala de puntuación utilizada corresponde a 1 = no se ajusta; 2 = se ajusta parcialmente; 3 = se ajusta completamente (Aranguren y Hoszowski, 2017).

III. Marco metodológico

A. Tipo de estudio

El presente estudio corresponde a una investigación teórica de tipo bibliométrica con un diseño descriptivo-retrospectivo (Montero y León, 2007). Este tipo de investigación se caracteriza por explorar la configuración intelectual de un campo de conocimiento sin aportar datos empíricos originales, sino basándose en el análisis de trabajos elaborados previamente. Hoyos (2010) explica que este tipo de investigación se corresponde con un procedimiento científico cuyo objetivo es lograr un conocimiento crítico sobre un tema de interés que, en este caso, corresponde al ámbito psicométrico de la definición de puntos de corte en pruebas de evaluación referidas a criterios.

El carácter descriptivo-retrospectivo del estudio está dado por el análisis de relaciones entre variables en una situación pasada, siendo las unidades de análisis objetos y no personas; que en este caso corresponden a los documentos o publicaciones analizadas (Montero y León, 2007). A partir de la identificación, selección y organización del material a analizar se amplía y profundiza en el tema abordado pero tomando en consideración una serie de elementos que permitan comprender de mejor manera la configuración intelectual del campo de conocimiento como base para el análisis crítico según el contexto de interés (Bernal, Martínez, Parra y Jiménez, 2015).

Angarita (2014) explica que los estudios bibliométricos no buscan únicamente conocer más acerca de un tema en particular, sino que buscan caracterizar el proceso o producción científica dentro de la cual se enmarca dicho tema. En el caso de la presente investigación se pretende que, además de realizar dicha exploración y análisis del campo de conocimiento del establecimiento de puntos de corte en pruebas de evaluación referidas a criterios, sea posible generar una propuesta de aplicación de una estrategia metodológica pertinente para el contexto de evaluación de estudiantes de Primaria en el ámbito educativo costarricense.

La revisión de la literatura dentro de este diseño de investigación se lleva a cabo con importante rigor metodológico con el fin de que las inferencias y las relaciones conceptuales que se identifiquen puedan, de manera confiable, dar cuenta del saber acumulado en el campo de conocimiento (Bernal, Martínez, Parra y Jiménez, 2015). De esta manera, con el fin de emitir inferencias confiables en el desarrollo del presente TFM se

toman en consideración los criterios de rigurosidad científica desarrollados por Barbosa, Barbosa y Rodríguez (2013), los cuales se muestran en la **Tabla 4**.

Tabla 4. Criterios de rigurosidad científica tomados en cuenta en el estudio

Criterio	Descripción
Finalidad	<p><i>Representa el compromiso por establecer objetivos de investigación previos.</i></p> <p>En el caso de la presente investigación, la revisión creativa de la literatura se elabora partiendo de unos objetivos concretos de investigación orientados a la identificación de estrategias pertinentes para definir puntos de corte en pruebas de evaluación de competencias digitales en estudiantes.</p>
Coherencia	<p><i>Refiere a la posibilidad de contar con unidad interna en materia de fases, actividades y datos.</i></p> <p>Como se explicita en el presente apartado de metodología, el estudio de TFM se desarrolló siguiendo un hilo investigativo de revisión creativa de la literatura, el cual determinó los procedimientos a desarrollar en coherencia con los objetivos de investigación que se plantearon desde el diseño de investigación.</p>
Fidelidad	<p><i>Alude a un respaldo en materia de recolección y transcripción.</i></p> <p>Durante el desarrollo del TFM se respaldaron las búsquedas de información realizadas, así como los diferentes tipos de análisis desarrollados. Todo el material se encuentra debidamente almacenado y disponible para cualquier consulta o análisis posterior que se requiera.</p>
Integración	<p><i>Implica articulación y evaluación global del proceso de investigación.</i></p> <p>En el sentido de la elaboración del análisis de la información, este criterio corresponde a las valoraciones realizadas en el marco de cada uno de los objetivos específicos de investigación y a la manera en la que éstos se vinculan para responder al objetivo general del estudio.</p>
Comprensión	<p><i>Se traduce en el favorecimiento de la construcción teórica sobre el objeto de estudio.</i></p> <p>El diseño de investigación bibliométrico descriptivo-retrospectivo buscó que a partir de la comprensión de la configuración intelectual del campo de conocimiento, se llevara a cabo un análisis crítico de las diferentes estrategias de definición de puntos de corte. Esta comprensión a su vez, sería útil para generar una propuesta de aplicación pertinente para el contexto de las pruebas de evaluación de estudiantes del PRONIE MEP-FOD en Costa Rica.</p>

(Fuente: Elaboración propia a partir del modelo de Barbosa, Barbosa y Rodríguez, 2013)

B. Muestra

Dentro de la metodología de los estudios bibliométricos, los documentos a analizar corresponden a la muestra del estudio y la unidad de análisis corresponde al campo de conocimiento explorado (Angarita, 2014). En el caso del presente TFM, los documentos a analizar se obtuvieron de la base de datos *ISI Web of Science* (en adelante WOS) del paquete *ISI Web of Knowledge* (de la empresa Thomson Reuters). La selección de esta base de datos responde a que, de acuerdo con Cortés (2008), se trata de una de las de mayor presencia científica y académica a nivel internacional.

La WOS únicamente indexa producción científica de revistas arbitradas y, según datos de Thomson Reuters (citado en Cortés, 2008), está compuesta por diferentes índices que, en este caso, constituyen el universo muestral del estudio (Véase el detalle en la **Tabla 5**).

Dentro de las limitaciones que supone una base de datos de este tipo se encuentra que es de pago, por lo que es importante aclarar que se accedió a ella gracias a los servicios brindados por el sistema de bibliotecas de la Universidad de Valladolid.

Tabla 5. Principales componentes del *ISI Web of Science* que integran el universo muestral del TFM

Componente	Periodicidad	Descripción
Science Citation Index Expanded	Con información desde el año 1900	Contiene información de más de 7,100 revistas académicas de alrededor de 150 disciplinas de ciencia y tecnología.
Social Sciences Citation Index	Con información desde 1956	Contiene información de más de 2,100 revistas académicas de cerca de 50 disciplinas dentro de las Ciencias Sociales y algunos materiales seleccionados de 3,500 revistas líderes en ciencia y tecnología.
Arts & Humanities Citation Index	Con información desde 1975	Indexa más de 1,200 revistas académicas en las artes y las humanidades y otros artículos seleccionados de más de 6,000 revistas de ciencia y tecnología).
Conference Proceedings Citation Index	Con información desde 1990	Indexa más de 110,000 actas de congresos publicadas en revistas o en libros, en alrededor de 250 disciplinas científicas.

(Fuente: Thomson Reuters citado en Cortés, 2008).

Como parte de la estrategia de selección de la muestra para el estudio, se definieron los siguientes criterios de inclusión:

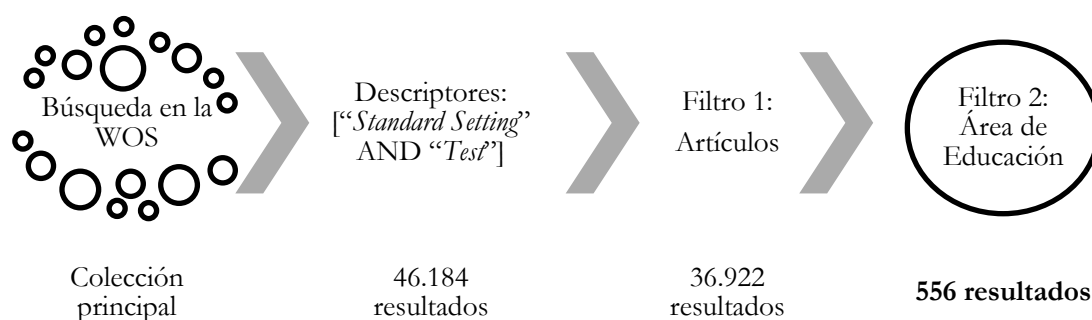
1. Estar redactado en formato de artículo científico.
2. Estar indexado en el *ISI Web of Science*.
3. Responder a los descriptores [*“Standard Setting”* AND *“Test”*] OR [*“Standard Setting”* AND *Cut scores*].
4. Corresponder al ámbito de la educación.

Partiendo de los criterios de selección anteriormente descritos, se llevó a cabo la primera búsqueda de información en la WOS, utilizando los descriptores mencionados. De esta primera búsqueda se recuperó un total de 556 fuentes de información pertinentes. En la **Tabla 6** se muestra el diagrama del procedimiento llevado a cabo:

Tabla 6. Procedimiento para la primera búsqueda de información en la WOS

Fecha de la búsqueda: 28 de marzo del 2018.

Procedimiento:



Script ejecutado:

Tema: (standard setting) AND **Tema:** (tests)

Refinado por: Tipos de documento: (ARTICLE) AND **Categorías de Web of Science:** (EDUCATION EDUCATIONAL RESEARCH OR EDUCATION SCIENTIFIC DISCIPLINES)

Índices=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC Período de tiempo=Todos los años

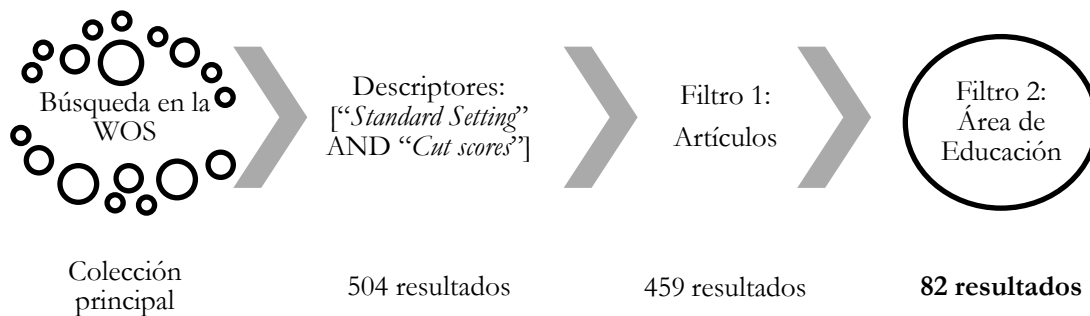
(Fuente: Elaboración propia a partir de las búsquedas de información realizadas en la WOS).

Posteriormente, se ejecutó la segunda búsqueda de información utilizando el otro set de descriptores seleccionados para aproximar el tema. En este caso, se recuperaron 82 fuentes de información más. En la **Tabla 7** muestra el procedimiento llevado a cabo:

Tabla 7. Procedimiento para la segunda búsqueda de información en la WOS

Fecha de la búsqueda: 28 de marzo del 2018.

Procedimiento:



Script ejecutado:

Tema: (standard setting) *AND* **Tema:** (cut scores)

Refinado por: Tipos de documento: (ARTICLE) *AND* **Categorías de Web of Science:**

(PSYCHOLOGY EDUCATIONAL OR EDUCATION EDUCATIONAL RESEARCH OR EDUCATION SCIENTIFIC DISCIPLINES OR EDUCATION SPECIAL)

Índices=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC Período de tiempo=Todos los años

(Fuente: Elaboración propia a partir de las búsquedas de información realizadas en la WOS).

Como resultado de ambas búsquedas se contó con un total de 638 artículos científicos. Sin embargo, al descartar aquellos que se repitieron como resultado de ambos procesos de búsqueda, la muestra final quedó conformada por un total de **601 artículos científicos** que cumplieron con los criterios de inclusión definidos para la investigación.

C. Procedimiento de recolección y análisis de los datos

Para la definición de las fases que guiarían el desarrollo del estudio, se tomó como base los modelos de Aria y Cuccurullo (2017) y Hoyos (2010) para el desarrollo de estudios que involucran la estrategia de *Science Mapping* dentro de un campo de conocimiento. Como se observa en la **Figura 15**, se establecieron cuatro fases para el desarrollo del estudio (Véase el cronograma detallado del desarrollo del TFM en **Anexo 3**):

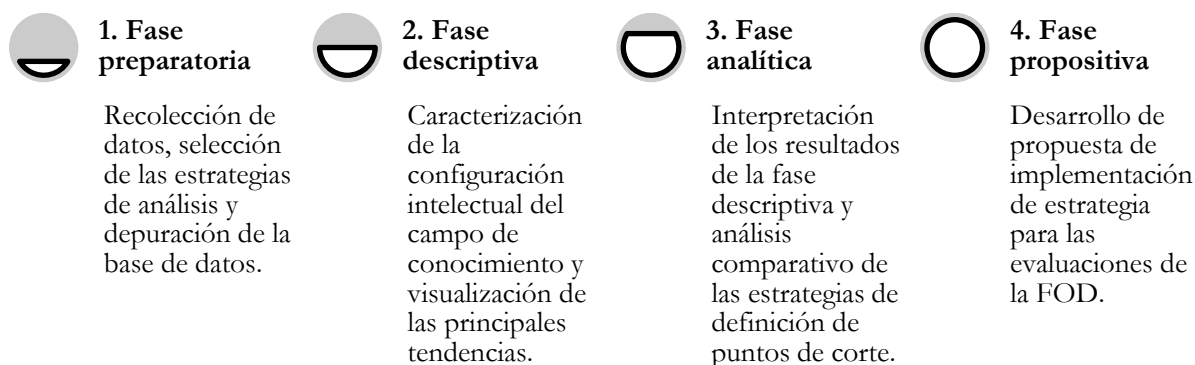


Figura 15. Síntesis del procedimiento de recolección y análisis de datos llevado a cabo para el desarrollo del TFM (Fuente: elaboración propia a partir de la definición de la metodología).

1. Fase preparatoria.

Se llevó a cabo la definición de los criterios de inclusión para seleccionar las fuentes de información a analizar. Posteriormente, se llevó a cabo la recolección de la información desde la WOS, siguiendo la estrategia de filtrado según los criterios establecidos. Dentro de esta fase se llevó a cabo la preparación de la base de datos a emplear, que implicó realizar un proceso de depuración y unificación de la información disponible (por ejemplo, de los nombres de los autores y otros ajustes), según los requerimientos de formato establecidos por las herramientas de análisis seleccionadas.

Dentro de esta fase se dio la preparación del *script* para ejecutar el análisis global de la base de datos, así como la definición de los parámetros y aspectos a considerar en dicho análisis. De esta manera, la fase preparatoria incluyó los procesos de recuperación, filtrado y pre-procesado de la base de datos para el estudio, así como la definición de las estrategias y las herramientas de análisis a emplear.

2. Fase descriptiva

Dentro de esta fase se inició el análisis de la información utilizando la herramienta de software R en su versión 3.4.4 (Véase *R Core Team*, 2018); específicamente, se empleó una herramienta útil para el análisis bibliométrico de tipo *Science Mapping*, la cual fue desarrollada por Aria y Cuccurullo (2017) que permite realizar análisis estructurados y caracterizaciones generales de amplios volúmenes de información científica mediante el software R. Para el desarrollo de la estrategia de *Science Mapping*, se empleó la herramienta *VOSViewer*, con el objetivo de visualizar y extraer las redes de nodos o mapas propios de la configuración intelectual del campo de conocimiento. Dicho análisis se llevó a cabo a partir de la definición de variables de interés como la conformación conceptual y semántica del campo de conocimiento, los autores clave, las tendencias de citación entre autores y los principales referentes por espacio geográfico, entre otros.

3. Fase analítica

En un primer momento, se llevó a cabo una interpretación de los resultados obtenidos en la fase descriptiva, con lo cual se caracterizó el campo de conocimiento en estudio. En un segundo momento, se realizó una aproximación más en profundidad dentro de un subconjunto de datos para analizar las principales estrategias de definición de puntos de corte identificadas y disponibles en la literatura. Mediante un método analítico que implicó “separar un todo de forma intelectual o material en sus partes individuales para hacer un estudio pormenorizado de cada una de ellas” (García, 2015, p. 78), se empleó una estrategia comparativa, cuyo eje estuvo orientado a la caracterización de cada estrategia según criterios de interés definidos.

Este análisis derivó en una valoración del procedimiento, las fortalezas y las debilidades de cada una de las estrategias identificadas. Esta fue la base para realizar una valoración de la pertinencia de cada una de ellas a la luz del contexto de las pruebas de evaluación de estudiantes que se desarrollan en la FOD. Cabe mencionar que la sistematización y el análisis de esta información se llevó a cabo mediante una estrategia de matrices de comparación y categorización, procesadas en formato de Excel.

4. Fase propositiva

Posterior a una validación de los resultados de la fase analítica con el equipo de trabajo de la Unidad de Evaluación de la FOD, en esta última fase se realizó una propuesta de implementación de la estrategia de definición de puntos de corte que se consideró más pertinente para el contexto evaluativo del PRONIE MEP-FOD. Este proceso implicó la derivación de pautas para la planificación de una metodología a seguir, así como la proyección de requerimientos por ejemplo en cuanto a elaboración de materiales y otros aspectos clave para su ejecución, durante la segunda mitad del año 2018 en la FOD.

IV. Resultados

A. Configuración intelectual del campo de conocimiento

En el presente apartado se muestran los resultados del análisis de *Science Mapping* elaborado para explorar la configuración intelectual del campo de conocimiento de la definición de puntos de corte en pruebas de evaluación referidas a criterios. Es importante enfatizar que estos resultados, si bien fueron la base para el desarrollo de la propuesta de implementación de una estrategia de definición de puntos de corte para la FOD, constituyen en sí mismos un importante aporte del TFM al campo de conocimiento en estudio.

Descripción del conjunto de datos

Para la caracterización del conjunto de datos se ejecutaron en dos momentos los análisis bibliométricos propuestos por Aria y Cuccurullo (2017; véase Script de *Bibliometrix* ejecutado en R en **Anexo 4**). Primero se analizó el conjunto de datos de manera exploratoria al momento de obtener el conjunto de datos desde la WOS (Véase *Inicial* en **Tabla 8**). Posterior a la depuración de la base de datos (Véase *Preprocesado* en **Tabla 8**), se ejecutaron de nuevo los análisis⁴. Como se observa, se trata de un conjunto de 601 artículos de 179 revistas científicas, cuyo rango de años de publicación está comprendido desde 1976 hasta el 2018. Además, se identifica la participación de al menos 1685 autores.

Al efectuar un análisis más detallado de los años de publicación se identifica que antes de la década de los 90 la frecuencia de publicación no superaba un artículo por año (Véase el detalle de las publicaciones por año en **Anexo 5: Tabla 1**). Posteriormente, entre los años 1990 y 2000, la producción aumentó a un promedio de 10 artículos por año, cifra que se duplicó en el periodo comprendido entre los años 2001 y 2010, en donde se tuvo un promedio de 20 publicaciones por año.

Es importante destacar que este campo de conocimiento parece haber tenido su mayor desarrollo en los años más recientes, ya que en lo que va de la década actual (desde el año 2011 hasta el año 2018) la producción científica ha llegado a alcanzar cerca de los 40 artículos por año. Así, por ejemplo, durante el año 2017 se alcanzó un tope de 61

⁴ La comparación entre los datos del análisis *inicial* y el *pre-procesado* se muestran únicamente en la **Tabla 7** para efectos de comparar los resultados de la depuración de la base de datos. Sin embargo, en los análisis y tablas posteriores (incluidos Anexos) únicamente se muestran los datos depurados.

publicaciones científicas, siendo el promedio general de crecimiento del campo de conocimiento de un 7,35% por año.

Tabla 8. Información principal del conjunto de datos

	Inicial	Preprocesado
Artículos	601	601
Fuentes (Revistas, libros, etc.)	179	179
Palabras clave Plus (ID)	985	985
Palabras clave del autor (DE)	1441	1441
Periodo	1976 - 2018	1976 - 2018
Promedio de citas por artículo	14,77	14,77
Autores	1717	1685
Apariciones del autor	1952	1952
Autores de artículos de autor único	95	91
Autores de artículos de varios autores	1622	1594
Artículos por autor	0,35	0,36
Autores por artículo	2,86	2,80
Co-Autores por artículo	3,25	3,25
Índice de colaboración	3,3	3,24

(Fuente: Traducción libre de salida de datos de *Bibliometrix* de R).

Los datos sobre la autoría de los documentos sugieren que, en la mayoría de los casos, se han realizado las publicaciones en conjunto con otros autores y no de manera individual. Lo anterior se confirma con el hecho de que se cuenta con un promedio de 0,36 artículos por autor, es decir, que el promedio de autores por artículo es cercano a 3 (concretamente 2,80). Pese a esto, algunos de los autores tienen altas tasas de productividad, como es el caso de Van del Vleutenc, quien se coloca como el autor más productivo dentro del conjunto de datos, con un registro de al menos 20 artículos en los que figura como primer autor (Véase el detalle de autores más productivos en **Anexo 5: Tabla 2**).

El análisis de los autores más productivos dentro del campo de conocimiento también permite ubicar los países de donde provienen dichos autores, con lo cual es posible generar un ranking de los más productivos dentro conjunto de datos estudiado (Véase detalle de países más productivos en **Anexo 5, Tabla 3**). Al respecto, en la **Figura 16** se destaca Estados Unidos como el país con mayor cantidad de publicaciones dentro del conjunto de datos (310 publicaciones, correspondientes al 51,08% del total que conforma la muestra). Adicionalmente, es posible identificar que este país se ha citado en al menos 5.231 oportunidades, lo cual equivale a un promedio de citación de 16,87 artículos (Véase detalle de las citaciones por país en **Anexo 5: Tabla 4**).

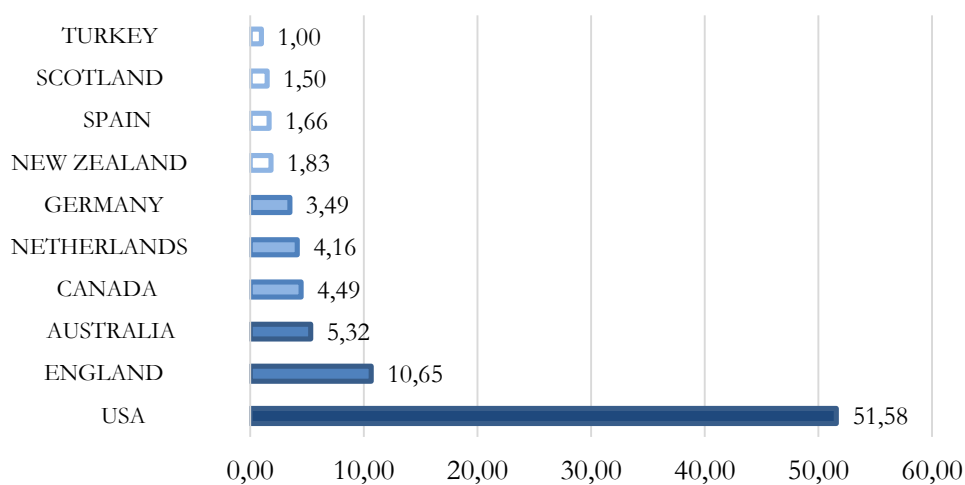


Figura 16. Porcentaje de aporte por país al campo de conocimiento estudiado (n=601 artículos) (Fuente: Elaboración propia a partir del conjunto de datos).

Algunos de los artículos que conforman la base de datos cuentan con altos índices de citación, es decir, que son importantes referentes dentro del campo de conocimiento. De esta manera, destaca el caso del artículo [ANDERSON JR; CORBETT AT; KOEDINGER KR; PELLETIER R, (1995), J. LEARN. SCI]⁵, que ha sido citado en al menos 642 ocasiones en los otros artículos que conforman el conjunto de datos (Véase detalle de artículos más citados en **Anexo 5: Tabla 5**).

Al analizar en detalle el contenido de este artículo, se observa que fue publicado en el año 1995 por el *Journal of the Learning Sciences* y se trata de una revisión de las publicaciones

⁵ Anderson, J., Corbett, A. Koedinger, K. & Pelletier, R. (1995). Cognitive Tutors: Lessons Learned. *Journal of the Learning Sciences*, 4(2), 167-207.

realizadas en los últimos 10 años sobre el desarrollo de tutores computarizados para apoyar el aprendizaje de las matemáticas de los estudiantes. Considerando que el desarrollo de este tipo de apoyo computarizado requiere de la definición de sistemas de evaluación que permitan identificar un nivel de competencia por parte del aprendiz, se encuentra coherencia en cuanto al peso que tiene este artículo dentro del conjunto de datos.

Otros artículos que destacan dentro de los más citados son [HELLER P; HOLLABAUGH M, (1992), AM. J. PHYS.]⁶, con un total 238 citaciones y el de [RETHANS JJ; NORCINI JJ; BARON-MALDONADO M; BLACKMORE D; JOLLY B; LADUCA T; LEW S; PAGE GG; SOUTHGATE L, (2002), MED. EDUC.]⁷, con un total de 189 citaciones. En ambos casos, se trata de artículos relacionados con la evaluación de habilidades. Sin embargo, el primero, aborda el tema de la evaluación de habilidades de resolución de problemas dentro del ámbito de la educación, mientras que el segundo, está más orientado a la diferenciación entre los modelos de medición basados en competencias y los basados en desempeños, en evaluaciones desarrolladas dentro del ámbito de la medicina.

Como parte de la caracterización del conjunto de datos, es posible identificar otras referencias clave que no son necesariamente artículos científicos, pero que igualmente tienen un importante peso dentro de la muestra (Véase detalle de referencias más frecuentemente citadas en **Anexo 5: Tabla 6**). En este caso, se destacan las referencias que se muestran en la **Tabla 9**, las cuales corresponden a dos libros y un artículo.

⁶ Heller, P. & Hollabaugh, M. (1992). Teaching problem-solving through cooperative grouping. Designing problems and structuring groups. *American Journal of Physics*, 60(7), 637-644. DOI: 10.1119/1.17118.

⁷ Rethans, J., Norcini, J., Baron-Maldonado, M., Blackmore, D., Jolly, B., LaDuca, T., Lew, S., Page, G., Southgate, L. (2002). The relationship between competence and performance: implications for assessing practice performance. *Medical Education*, 36(10), 901-909. DOI: 10.1046/j.1365-2923.2002.01316.x

Tabla 9. Otras referencias frecuentemente citadas dentro del conjunto de datos

Información de <i>R</i>	Frec	Referencia completa
[ANGOFF W H, 1971, ED MEASUREMENT, P508]	83	Libro: Angoff, W.H. (1971). <i>Scales, norms and equivalent scores. Educational Measurements.</i> Washington, DC: American Council on Education
[NORCINI JJ, 2003, MED EDUC, V37, P464, DOI 101046/J1365-2923200301495X]	30	Artículo: Norcini, J. (2003). Setting standards on educational tests. <i>Medical Education</i> , 37(5), pp. 464-469.
[CIZEK G J, 2007, STANDARD SETTING GUI]	29	Libro: Cizek, G. & Bunch, M. (2007). <i>Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests.</i> Thousand Oaks, CA: Sage Publications Ltd.

(Fuente: Traducción libre de la salida de datos del paquete *Bibliometrix* de R).

Análisis de redes y mapas de tendencias del conjunto de datos

Partiendo de la estrategia de *Scienc Mapping*, como primera línea de interés se generó un análisis de la semántica del conjunto de datos que, de acuerdo con Aria y Cucurullo (2017) es útil para visualizar la estructura cognitiva de un campo de conocimiento. En primer lugar, se realizó un análisis de densidad de la co-ocurrencia (*co-word*) de palabras clave (*keywords*) identificadas en el conjunto de datos. Para ello, se definieron como parámetros de análisis:

- 1. Utilizar las palabras clave definidas por los autores y las asignadas por la base de datos.** En este caso, entraron en el análisis los 601 títulos con las respectivas palabras clave asignadas por los autores [*Authors Keywords*] y por la base de datos, al momento de indexar cada artículo [*Keywords-Plus*].
- 2. Tomar en cuenta todas las palabras clave identificadas (*full counting*).** Se identificó un total de 2180 palabras clave como descriptores de los artículos que conforman el conjunto de datos analizado.
- 3. Utilizar en la red aquellas palabras clave que aparecieran mínimo 5 veces en el conjunto de datos.** Se refinó la búsqueda con este parámetro para identificar las palabras clave que tuvieran un mayor peso dentro del conjunto de datos. De esta manera, se redujo a 102 la cantidad palabras clave identificadas.

Como se muestra en la **Figura 17**, destacan cuatro palabras clave como principales referentes semánticos en el conjunto de datos, las cuales corresponden a:

1. [EDUCATION], 2. [PERFORMANCE], 3. [ASSESSMENT] y 4. [STANDARD SETTING].

En el caso de las dos primeras palabras clave con mayor densidad, se identificó que su mayor grado de co-ocurrencia proviene de las palabras clave definidas por la base de datos al momento de indexar los artículos (frecuencia total = 81 y 53 artículos, respectivamente)⁸, mientras que las otras dos palabras, tuvieron más co-ocurrencia por parte de las palabras clave que definen los autores (frecuencia total = 41 y 39 artículos, respectivamente). Adicionalmente, dentro del conjunto de palabras clave destacan otras como zonas de alta densidad, es decir, que también fueron utilizadas frecuentemente, como son las relacionadas con *habilidades*, *currículum* y *estándares de desempeño*, entre otras.

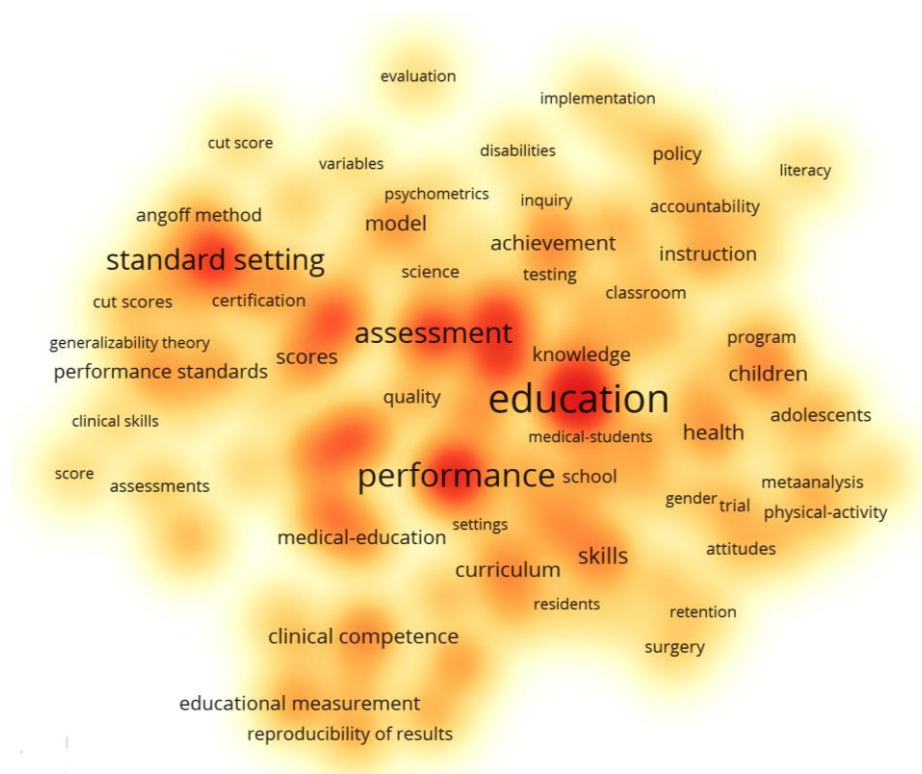


Figura 17. Análisis de co-ocurrencia de palabras clave definidas por los autores [*authors keywords*] y asignadas por la base de datos [*keywordsplus*] (n= 601 artículos, 2180 palabras clave identificadas, ≥ 5

⁸ El detalle de las palabras clave más frecuentemente utilizadas en los artículos que conforman el conjunto de datos se puede observar en el **Anexo 5: Tabla 7**).

repeticiones = 99 palabras representadas en la red). (Fuente: Elaboración propia a partir del conjunto de datos).

En segundo lugar, para ampliar la caracterización de la semántica vinculada al campo de conocimiento, se llevó a cabo un análisis de co-ocurrencia de los términos empleados en otras partes del cuerpo de los artículos. Para ello, se definieron como parámetros:

- 1. Utilizar los términos disponibles en el título y el resumen de los artículos.** En este caso, entraron en el análisis los 601 títulos y resúmenes correspondientes a cada uno de los artículos que conforman el conjunto de datos.
- 2. Tomar en cuenta todos los términos identificados (*full counting*).** Se identificó un total de 14.246 términos en el título y resumen de cada uno de los artículos que conforman el conjunto de datos.
- 3. Utilizar en la red aquellos términos que aparecieran mínimo 25 veces en el conjunto de datos.** Pese a que el programa sugiere considerar aquellos que aparecen al menos 10 veces, se refinó la búsqueda con el fin de identificar los términos que tuvieran un mayor peso dentro del conjunto de datos. De esta manera, se redujo a 178 la cantidad de términos que cumplían con este parámetro.
- 4. Visualizar la red utilizando el 70% de los términos que cumplieran con los parámetros establecidos.** En este caso, dicho porcentaje correspondió a un total de 125 términos. Como valor adicional, se definió visualizar un máximo de 500 líneas de asociación entre los términos identificados.

Como se observa en la **Figura 18**, dentro del conjunto de datos se identificó una red semántica compuesta por tres clústeres o subgrupos que representan las áreas temáticas en las que los términos identificados tienden a estar más relacionados. Al visualizar en detalle la red generada, se observa que el primer clúster (representado con el color rojo) se agrupa principalmente en torno al concepto [*Student*] o estudiante y está vinculado a un total de 51 términos relacionados con temas educativos como son: aprendizaje, escuela, maestro y logro académico, entre otros. Este subconjunto de la red muestra que del total de 601 artículos que conforman la base de datos destaca un importante grupo que tiene que ver con el ámbito de la educación, lo cual es coherente con los descriptores y filtros empleados para llevar a cabo la búsqueda.

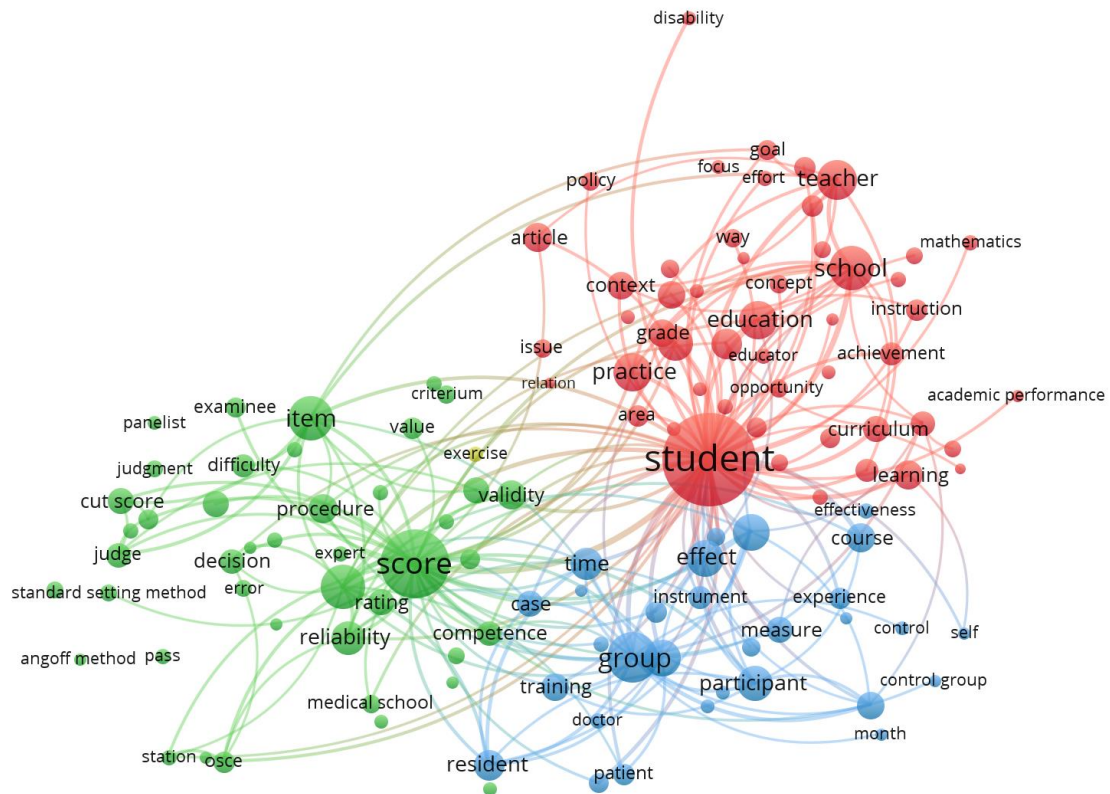


Figura 18. Análisis de co-ocurrencia de términos del título y resumen de los artículos (Min. Strength = 100, Max. Lines = 500, n= 601 artículos, 14.246 conceptos, ≥ 25 repeticiones, 70% representados en la red). (Fuente: Elaboración propia a partir del conjunto de datos).

El segundo clúster identificado (representado con color verde) se agrupa principalmente en torno al concepto *[Score]* o puntuación y está relacionado con 43 conceptos cuya temática está más orientada a aspectos metodológicos de la definición de puntos de corte. En este caso, se destacan conceptos como: ítem, puntos de corte, definición de estándares, validez, confiabilidad, juicios y expertos, entre otros. Este clúster también resulta coherente con los términos de búsqueda empleados para definir el conjunto de datos, los cuales estaban orientados específicamente a aspectos relacionados con las metodologías o estrategias psicométricas útiles para establecer puntos de corte en pruebas de evaluación.

En el caso del tercer clúster (representado con el color azul), se encontró que está compuesto por 30 términos y mostró una tendencia a estar más vinculado con el campo de la metodología de evaluación pero dentro del área de evaluación en medicina. Agrupados alrededor del concepto *[Group]* o grupo, en este clúster destacan términos como: participantes, medida, instrumento, paciente, residente y grupo control, entre otros. Pese a que este subtema no estaba contemplado dentro de las líneas de interés definidas en el

TFM, se considera un hallazgo interesante, ya que sugiere que las metodologías de definición de puntos de corte también son utilizadas en la definición de parámetros de certificación de competencias o de licencias profesionales en el campo de la salud.

Por otra parte, la red conceptual se analizó en función de la temporalidad con la que han ido emergiendo los términos. En la **Figura 19** es posible visualizar dicha evolución conceptual, en donde se identifica la tendencia de que los primeros términos emergentes estuvieron más vinculados con la metodología, es decir, con el segundo clúster.

Posteriormente, las publicaciones generadas alrededor de estos términos se empezaron a enfocar en las áreas de conocimiento específicas, identificándose primero los términos vinculados al ámbito de la educación y el aprendizaje de los estudiantes (primer clúster) y, más recientemente, los relacionados con áreas de evaluación en salud (tercer clúster).

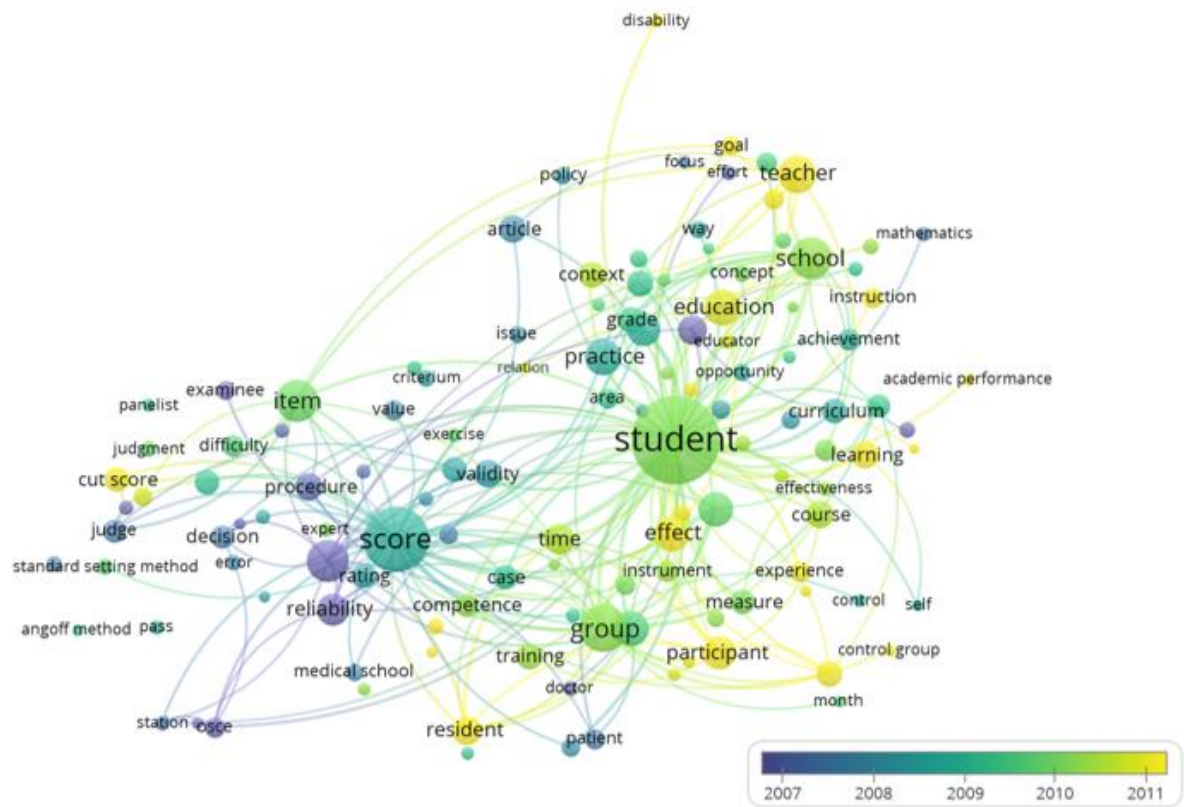


Figura 19. Evolución temporal de los principales términos de las publicaciones en el campo de conocimiento (Min. Strength = 100, Max. Lines = 500, n= 601 artículos, 14.246 conceptos, \geq 25 repeticiones, 70% representados en la red). (Fuente: Elaboración propia a partir del conjunto de datos).

La segunda línea de análisis explorada fue la tendencia de citación por región geográfica, con el objetivo de identificar aquellos países que parecen estar determinando de manera más importante los avances en el campo de conocimiento estudiado. En primer lugar, se realizó un análisis de co-citación (*co-citation*) entre los países con el objetivo de identificar los principales referentes y la manera en la que se vinculan las regiones. Para ello, se definieron como parámetros de análisis:

1. **Tomar en cuenta todos los países citados en cada artículo.** Del total de 601 artículos que componen el conjunto de datos se tuvo registro de 51 países involucrados en las publicaciones.
2. **Visualizar en la red aquellos países que tuvieran un mínimo de tres artículos asociados.** De esta manera, se incluyeron en la visualización de la red un total de 27 países que cumplían dicha condición.

En la **Figura 20**, es posible visualizar la tendencia identificada, en la que se destaca Estados Unidos como principal referente citado en las diferentes publicaciones, lo cual concuerda con lo observado en la descripción general del conjunto de datos, donde este país se identificó como el más productivo. La tendencia de citación de este país como referente, muestra que tiene relación con países como Inglaterra, Holanda, Canadá y Australia. En segundo lugar, se identifica relación menor con otros países europeos como Alemania, España, Escocia y Bélgica.

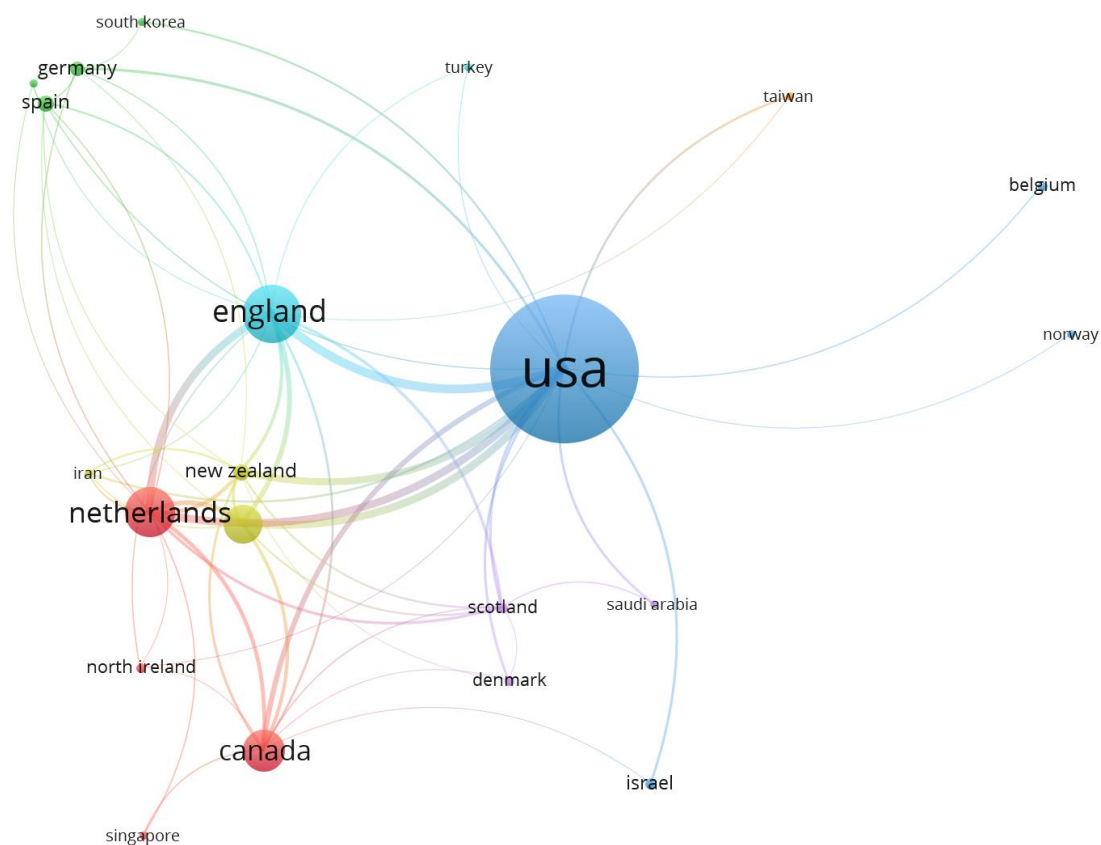


Figura 20. Análisis de citación entre países dentro del conjunto de datos (n= 601 artículos, Max. countries per document= 51, Min. documents per country= 3). (Fuente: Elaboración propia a partir del conjunto de datos).

En este análisis también se exploró la evolución temporal y, como se muestra en la **Figura 21**, los primeros países que empezaron a ser referentes en este campo fueron Holanda, Canadá, Nueva Zelanda y Escocia. Adicionalmente, se observa que los dos países que tienden a ser más citados en la actualidad (Estados Unidos e Inglaterra), empezaron a ser citados de manera más reciente; lo cual sugiere se colocaron como referentes en los últimos años. Asimismo, es posible visualizar que España, Alemania, Singapur e Irán se han incorporado de manera más reciente al campo de conocimiento.

La utilidad de esta red se relaciona con la posibilidad de visualizar qué países o regiones podrían tener mayor experiencia en el campo de conocimiento estudiado, que en este caso parecen ser Europa y Estados Unidos. Un aspecto que llama la atención es que no se visualizan dentro del conjunto de datos referentes de peso que pertenezcan a la región latinoamericana. Esto sugiere que el tema aún no se ha desarrollado de manera consolidada en tal región, a pesar de que se cuenta con algunos esfuerzos identificados.

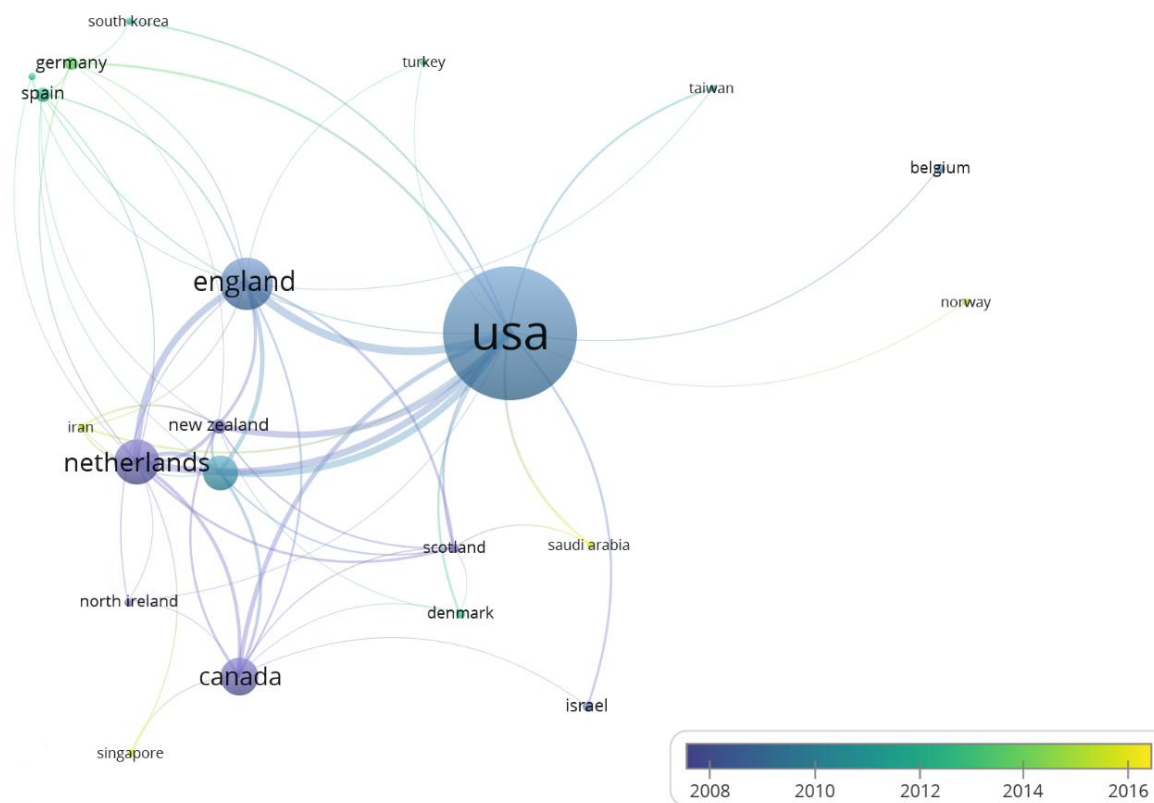


Figura 21. Evolución temporal de la citación entre países dentro del conjunto de datos (n= 601 artículos, Max. countries per document= 51, Min. documents per country= 3). (Fuente: Elaboración propia a partir del conjunto de datos).

En general se encuentra que la configuración intelectual del campo de conocimiento muestra que se trata de un área con varias décadas de desarrollo, pero en la que su mayor volumen de aportaciones viene dado por los trabajos de los últimos años. Adicionalmente, la semántica identificada guarda coherencia con el área de conocimiento dentro de la cual se enmarca el estudio.

Este hallazgo da solidez a la selección de los descriptores empleados para llevar a cabo las búsquedas de información en la WOS que fueron la base del estudio. Asimismo, los resultados permiten evidenciar las principales tendencias en cuanto a autores y países más influyentes dentro del campo de conocimiento; lo cual se considera oportuno dentro del objetivo al que responden estos resultados en el TFM.

B. Análisis de las estrategias identificadas según pertinencia para la FOD

En concordancia con el objetivo de realizar una comparación de las principales estrategias de definición de puntos de corte identificadas en la literatura, se llevó a cabo una exploración de sus principales características como base para el análisis de pertinencia en función del contexto de evaluación del PRONIE MEP-FOD. Los criterios de comparación fueron definidos en conjunto con la Unidad de Evaluación y se describen en la **Tabla 10**, dentro de los principales aspectos contemplados destacan la necesidad de que las estrategias permitieran derivar más de un punto de corte y que permitiera trabajar con ítems de selección única, entre otros.

Una vez definidos los criterios de comparación de las estrategias, el punto de partida para su identificación fue uno de los referentes claves identificados en la exploración de la configuración intelectual de campo de conocimiento, el cual corresponde al libro de Cizek y Bunch (2007). Adicionalmente, el análisis se complementó con cuatro fuentes primarias más (Aranguren y Hoszowski, 2017; Barrios y Cosculluela, 2013; Leyva, 2011 y Muñiz, 2018), así como con algunos otros referentes citados por dichos autores.

Tabla 10. Criterios de comparación de estrategias contemplados en el análisis

Criterio	Descripción
Clasificación	Se clasificaron en <i>Test centered</i> , <i>Examinee centered</i> y <i>Métodos de compromiso</i> .
Ámbito de uso	Correspondiente al ámbito y finalidad para los que se suelen aplicar las estrategias. Por ejemplo, en educación para evaluar estudiantes, etc.
Temporalidad	Entendido como el momento en el que es posible implementar la estrategia. Es decir, antes o después de contar con datos empíricos.
Tipo de ítem	Especifica si la estrategia admite únicamente ítems de selección única o si admite otro tipo de formatos como ítems de respuesta abierta, etc.
Cantidad de jueces	Corresponde a las sugerencias que se brindan en las diferentes estrategias sobre el número de jueces que se deberían contemplar.
Tipo de punto de corte	Refiere a si la estrategia permite derivar varios niveles de o si se trata de una estrategia en la que el resultado es un punto de corte binario (aprobado o no aprobado, por ejemplo).
Procedimiento	Detalla las premisas fundamentales sobre las que se basa, a la vez que deriva los pasos para su implementación. Se especifican consignas, procedimiento estadístico implicado y otras consideraciones útiles.
Ventajas y alcances	Sintetiza los aspectos que la literatura destaca como claves para colocar la estrategia como una más favorable.
Desventajas y limitaciones	Sintetiza los aspectos que la literatura destaca como claves para colocar la estrategia como una menos favorable.
Pertinencia para la FOD	Valoración global de la estrategia en función del contexto FOD. Se definieron tres categorías: <i>Pertinencia Baja (PB)</i> , <i>Media (PM)</i> y <i>Alta (PA)</i> .

Nota: la sistematización completa de esta información se realizó en una matriz comparativa en formato de Excel. Sin embargo, en el **Anexo 6** se incluye una ficha de síntesis de las ventajas y desventajas identificadas para cada estrategia.

Caracterización general de las estrategias identificadas

Se identificaron **17 estrategias** de definición de puntos de corte útiles para implementar en pruebas de evaluación referidas a criterios (14 estrategias principales y tres que corresponden a variaciones de una de ellas). En la **Figura 22** se muestra la distribución de dichas estrategias según la tipología de Barrios y Cosculluela (2013), según la cual se identificaron ocho⁹ estrategias centradas en la prueba o *Test Centered*, seis centradas en el examinado o *Examinee centered*, y tres del tipo *Métodos de compromiso*.

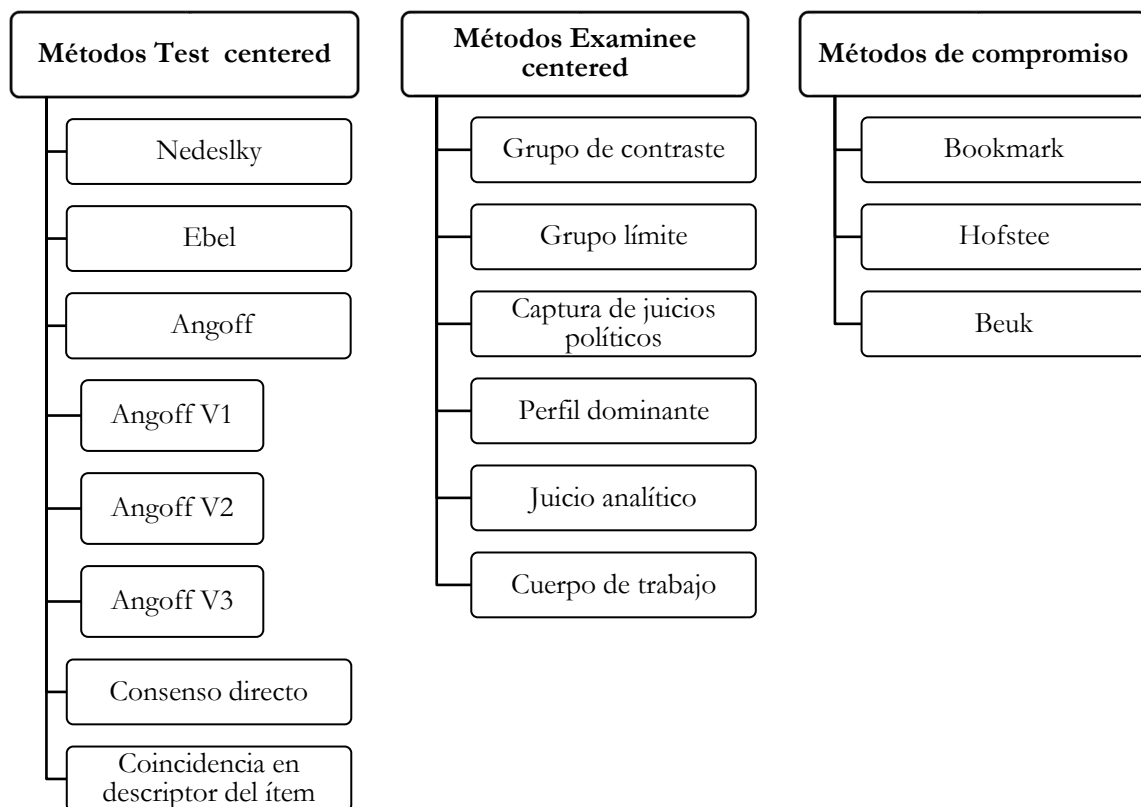


Figura 22. Clasificación de las estrategias de definición de puntos de corte para pruebas de evaluación referidas a criterios (Fuente: Elaboración propia a partir del conjunto de datos).

Al realizar la caracterización inicial, se encontró que la mayoría de estrategias se emplean en el ámbito de la educación (con excepción del método *Ebel*, que se emplea principalmente en salud). Algunas se utilizan con mayor regularidad en la emisión de certificaciones o asignación de licencias profesionales (como los métodos *Angoff*, *Consenso directo*, *Captura de juicios políticos*, *Perfil dominante*, *Juicio analítico* y *Beuk*) y otras se utilizan principalmente para

⁹ Se contabiliza como cuatro distintas estrategias el método Angoff con sus respectivas variaciones.

evaluar desempeño de estudiantes (como los métodos *Nedelsky*, *Grupos de contraste*, *Grupo límite*, *Cuerpo de trabajo*, *Bookmark* y *Hofstee* y las variaciones del *Angoff*).

Dentro de los criterios de comparación definidos destacan con mayor relevancia para el estudio el tipo de puntos de corte que se generan a partir de la implementación de la estrategia, así como el momento en el que se debe implementar. De esta manera, en la **Tabla 11** se muestra una organización de las estrategias en un plano que interpone ambos criterios, observándose que la mayoría requiere contar con datos empíricos para poder ser implementadas. Sin embargo, solo unas pocas permiten establecer puntos de corte que no sean dicotómicos.

Tabla 11. Distribución de estrategias según requerimiento de datos empíricos y tipo de punto de corte que genera su implementación

		Tipo de punto de corte	
		a. Dicotómico	b. Varios niveles
Datos empíricos	1. Requiere	Angoff V3 Grupos de contraste Grupo límite Captura de juicios políticos Perfil dominante Hofstee Beuk	Angoff V1 Angoff V2 Coincidencia por descriptor del ítem Juicio analítico Cuerpo de trabajo Bookmark
	2. No requiere	Nedelsky Ebel(*) Angoff Consenso directo(*)	

(*): Admiten modificación en la que se puede trabajar con datos empíricos.

(Fuente: elaboración propia a partir del conjunto de datos).

Partiendo del hecho de que en el contexto de evaluación de la FOD interesaría que la estrategia permita definir varios niveles de competencia, en este punto del análisis es posible reducir la cantidad de estrategias que podrían ser pertinentes para dicho contexto a aquellas que se encuentran ubicados en el cuadrante [b.1.] de la figura anterior.

Otro aspecto contemplado en la comparación de las estrategias corresponde al tipo de ítems con los cuales se puede trabajar. En este caso, se identifica que la mayoría trabaja con ítems de respuesta múltiple con selección única, es decir, con el tipo de ítem que se trabaja en las pruebas de evaluación de la FOD. Algunos otros métodos admiten otros formatos de ítem, lo cual se considera un valor agregado porque abre la posibilidad de implementarse en otros tipos de pruebas. Al analizar con detalle el tipo de ítem asociado a las estrategias que se van ubicando como más pertinentes para el contexto de la FOD (ubicadas en el cuadrante [b.1.]), se obtiene la distribución que se muestra en la **Tabla 12**:

Tabla 12. Tipos de ítem de las estrategias prioritarias para el contexto de la FOD

	Selección única	Respuesta abierta	Otros formatos
Angoff V1	✓	✓	
Angoff V2	✓	✓	
Coincidencia en descriptor del ítem	✓	✓	
Juicio analítico		✓	✓
Cuerpo de trabajo		✓	✓
Bookmark	✓	✓	

(Fuente: Elaboración propia a partir del conjunto de datos).

Al igual que en el caso de las características analizadas con anterioridad, este aspecto permitió reducir aún más las estrategias que podrían ser más pertinentes para el contexto de evaluación de la FOD al considerar que, por el momento, se trabaja sobre la base de una prueba cuyos ítems son de selección única. Partiendo de lo anterior, se descartan las estrategias de *Juicio analítico* y *Cuerpo de trabajo*, al no admitir dicho tipo de ítem.

En cuanto a la cantidad de jueces recomendados para la implementación de cada una de las estrategias identificadas, se encuentra una ausencia de recomendación concreta.

Únicamente se mencionan algunos ejemplos cuyas proporciones son variadas entre un autor y otro. Sin embargo, una distinción que se hace se relaciona con el hecho de que los jueces deban conocer directamente o no a los sujetos evaluados. Dicha condición sería

menos oportuna para el contexto de evaluación de la FOD porque se trata de un sistema de evaluación a gran escala en el que se suele trabajar con muestras representativas de gran volumen. Por lo cual, la posibilidad de reclutar como jueces a todos los docentes de los estudiantes evaluados a nivel nacional resultaría poco viable.

Dentro del subconjunto de estrategias identificadas como más pertinentes para las evaluaciones de estudiantes beneficiarios del PRONIE MEP-FOD, se identifica este requerimiento en la estrategia de *Coincidencia por descriptor del ítem*, con lo cual se excluye. De esta manera, la primera caracterización evidencia que de las 17 identificadas inicialmente, se reduce a un subgrupo de tres las que cumplen con los principales requerimientos del contexto de evaluación que se desarrolla desde la FOD. Concretamente, dichas estrategias corresponden a las variaciones 1 y 2 del método *Angoff* y a la estrategia de *Bookmark*.

Valoración de fortalezas y debilidades de las estrategias

La pertinencia de las estrategias de definición de puntos de corte se analizó en función de las características generales y de las fortalezas y debilidades identificadas en la literatura para cada una de éstas. Para ello, se generaron tres categorías que se muestran en la **Tabla 13**.

Tabla 13. Modelo de categorización de la pertinencia de las estrategias según el contexto de evaluación de la FOD

Pertinencia	Descriptor de la categoría
Pertinencia Baja (PB)	Presenta pocas características relacionadas con lo que se espera del modelo de definición de puntos de corte en el contexto de evaluación de la FOD. Puede mostrar importantes debilidades a nivel metodológico o poco respaldo en la literatura.
Pertinencia Media (PM)	Cumple con algunas de las características requeridas para el modelo de evaluación de la FOD pero cuenta con uno o dos elementos en los que no hay coincidencia. Puede tener algunas debilidades a nivel metodológico.
Pertinencia Alta (PA)	Sobresale con respecto a las coincidencias con lo que se espera de un modelo de definición de puntos de corte en el contexto evaluativo de la FOD. Se caracteriza por tener características metodológicas confiables y por estar respaldado en la literatura.

(Fuente: Elaboración propia a partir de la versión de Valdivia, 2012).

Cabe mencionar que este análisis fue sometido a un proceso de validación¹⁰ con el equipo de trabajo de la Unidad de Evaluación de la FOD en Costa Rica. Para ello, dos investigadores vinculados al proyecto (con diferentes áreas de especialidad), hicieron una categorización propia de cada una de las estrategias. Posteriormente, los resultados fueron discutidos en una puesta en común y, a partir de ello, se emitió la categorización final de cada una de las estrategias, según su pertinencia.

Como se muestra en la **Figura 23**, del total de 17 estrategias inicialmente identificadas, fue posible categorizar 10 como más distantes de los requerimientos del contexto de evaluación de la FOD, es decir, que se excluyeron por tener Pertinencia Baja (PA, en color rojo). Por su parte, cinco estrategias fueron valoradas como algo más cercanas a dicho contexto, es decir, que se categorizaron como con Pertinencia Media (PM, en color amarillo). Finalmente, dos de las 17 fueron las que resultaron ser más cercanas, es decir, que fueron valoradas como estrategias con Pertinencia Alta (PA, en color verde).

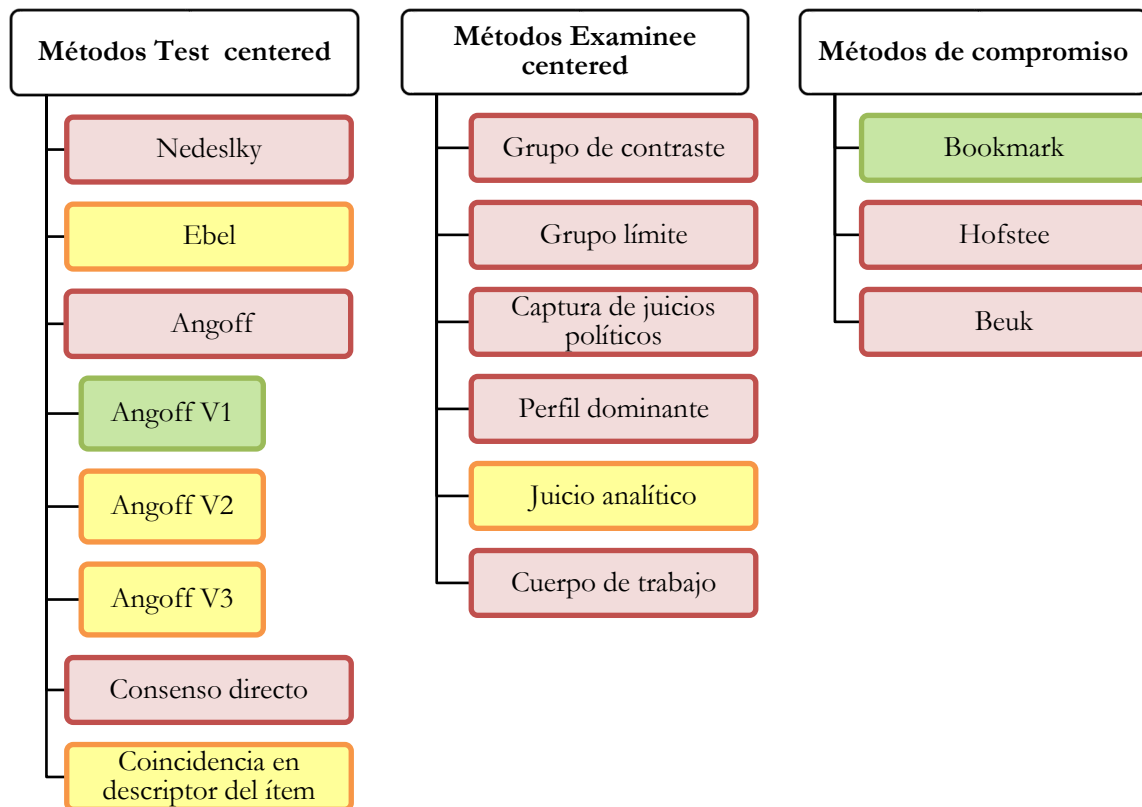


Figura 23. Categorización las estrategias según pertinencia para la FOD (Fuente: Elaboración propia a partir del conjunto de datos).

¹⁰ Véase síntesis de la sesión en **Anexo 7**.

1. Estrategias con Pertinencia baja (PB)

El análisis permitió descartar diez estrategias de las 17 inicialmente identificadas, es decir, que se ubicaron en esta categoría por no cumplir con algunas de las características fundamentalmente requeridas en el contexto de evaluación de la FOD. Como se observa en la **Tabla 14**, algunos criterios por los cuales se descartaron estas estrategias fueron el derivar puntos de corte dicotómicos y el mostrar algunas debilidades metodológicas.

Tabla 14. Síntesis de valoraciones para las estrategias categorizadas como Pertinencia Baja

Estrategia	Categoría de pertinencia: Baja.
Nedeslky	<ul style="list-style-type: none">• Orientada a brindar puntos de corte dicotómicos.• Genera los puntajes más bajos en los puntos de corte.
Angoff original	<ul style="list-style-type: none">• Orientada a brindar puntos de corte dicotómicos.• Identificada como una de las estrategias más débiles en confiabilidad.
Consenso directo	<ul style="list-style-type: none">• Orientada a brindar puntos de corte dicotómicos.
Grupos de contraste	<ul style="list-style-type: none">• Orientada a brindar puntos de corte dicotómicos.• Requiere que los jueces conozcan directamente a los evaluados.
Grupos límite	<ul style="list-style-type: none">• Orientada a brindar puntos de corte dicotómicos.• Requiere que los jueces conozcan directamente a los evaluados.
Juicios políticos	<ul style="list-style-type: none">• Orientada a brindar puntos de corte dicotómicos.• Identificada como una estrategia con limitaciones metodológicas. .
Perfil dominante	<ul style="list-style-type: none">• Orientada a brindar puntos de corte dicotómicos.• Requiere que los jueces conozcan directamente a los evaluados.• Identificada como una estrategia con limitaciones metodológicas.
Cuerpo de trabajo	<ul style="list-style-type: none">• Enfocada en el trabajo con otro tipo de formato de ítem.
Hofstee	<ul style="list-style-type: none">• Orientada a brindar puntos de corte dicotómicos.• Identificada como una estrategia con limitaciones metodológicas.• No se recomienda emplear como estrategia principal.
Beuk	<ul style="list-style-type: none">• Orientada a brindar puntos de corte dicotómicos.• No se recomienda emplear como estrategia principal.

(Fuente: Elaboración propia a partir del conjunto de datos).

2. Estrategias con Pertinencia Media (PM)

Dentro de esta categoría se ubicaron cinco de las 17 estrategias identificadas. Como se menciona en el descriptor de esta categoría, estas corresponden a las estrategias que cumplen con algunas de las características requeridas para el contexto de la FOD, pero que no coinciden en algún aspecto importante (Véase **Tabla 15**). Como ejemplo, destaca el caso de la variación 2 del método Angoff, que inicialmente cumplía con la mayoría de criterios requeridos pero cuyo análisis en detalle muestra que se utiliza principalmente para valorar ítems de respuesta abierta y, por ahora, este tipo de formatos no han sido prioritarios para la FOD.

Tabla 15. Síntesis de valoraciones para las estrategias categorizadas como Pertinencia Media

Estrategia	Categoría de pertinencia: Media.
Ebel	<ul style="list-style-type: none">• Permitiría comparar los criterios de dificultad y relevancia asignados por los jueces con los criterios bajo los que se construyeron los ítems.• No requiere que los jueces conozcan directamente a los evaluados.• Orientada a brindar puntos de corte dicotómicos.
Angoff 2	<ul style="list-style-type: none">• Orientada a brindar puntos de corte para varios niveles.• Enfocada principalmente en ítems de producción o respuesta abierta.
Angoff V3	<ul style="list-style-type: none">• Mejora el método Angoff original al incluir información empírica.• Orientada a brindar puntos de corte dicotómicos.
Coincidencia en descriptor del ítem	<ul style="list-style-type: none">• Orientada a brindar puntos de corte para varios niveles.• Requiere que los jueces conozcan directamente a los evaluados.
Juicio analítico	<ul style="list-style-type: none">• Orientada a brindar puntos de corte para varios niveles.• Requiere que los jueces conozcan directamente a los evaluados.• Implica una tarea altamente compleja para el juez.• No cuenta con suficiente documentación empírica.

(Fuente: Elaboración propia a partir del conjunto de datos).

3. Estrategias con Pertinencia Alta (PA)

Las estrategias *Angoff en su variación 1* y *Bookmark* sobresalieron como las dos más cercanas a los requerimientos propios del contexto de las evaluaciones de la FOD. En este caso, en la **Figura 24** y la **Figura 25** se detallan las ventajas y desventajas identificadas para cada una de las estrategias, aspectos que justificaron la valoración de pertinencia realizada.



- Reduce la variabilidad dentro del grupo de jueces al utilizar varias rondas de discusión para favorecer el consenso.
- Al incorporar información empírica, reduce la crítica que se hace al método *Angoff* original de estar basado en un criterio hipotético.



- Cuando se requiere definir más de un punto de corte se debe someter a los jueces a un mayor trabajo o demanda cognitiva.
- Requiere desarrollar una correcta conceptualización de los sujetos límite según los diferentes niveles requeridos, lo cual requiere más trabajo para los jueces.
- Su funcionamiento no ha sido ampliamente estudiado en otros formatos de ítem (como escalas Likert).

Figura 24. Síntesis de ventajas y desventajas identificadas en la estrategia *Angoff V1*. (Fuente: Elaboración propia a partir del conjunto de datos).



- Al estar basada en modelos de TRI, contempla parámetros de calidad, que minimizan errores de medida.
- Implica una tarea relativamente más sencilla para los jueces y con la que podrían estar más familiarizados.
- Los jueces tienen la posibilidad de profundizar su análisis centrándose en un rango más estrecho de ítems.
- Es uno de los pocos métodos para los cuales se cuenta con experiencias latinoamericanas (ver Aranguren y Hoszowski, 2017 y ENLACES, 2012).



- Requiere importante capacitación y práctica previas de los jueces para que comprendan la tarea.
- Existe la preocupación en torno a que los puntos de corte derivados son relativos al estar determinados por el grado de dificultad que la prueba tuvo para un conjunto determinado de sujetos. Sin embargo, esta limitación no es exclusiva del método sino que es generalizable a las estrategias que utilizan datos empíricos.
- Se requiere conocimiento estadístico en TRI y software como SPSS, Excel, Winsteps, etc. para llevar a cabo el análisis previo.

Figura 25. Síntesis de ventajas y desventajas identificadas en la estrategia *Bookmark*. (Fuente: Elaboración propia a partir del conjunto de datos).

Partiendo de la información descrita con anterioridad, se emitieron las siguientes valoraciones globales sobre ambas estrategias:

- **Estrategia Angoff V1: Pertinencia alta.**

Esta variación del método Angoff original viene a resolver las principales preocupaciones que se tenían con éste al estar basado en juicios elaborados sobre un grupo hipotético de estudiantes. En términos de pertinencia para los procesos de evaluación de la FOD, se valora que podría ser útil en tanto que se ha utilizado más en el ámbito de la educación, permite definir más de un punto de corte y admite la posibilidad de trabajar con ítems de selección única, así como con otros formatos.

La estrategia da valor a las discusiones entre jueces, lo cual coincide con el estilo de trabajo institucional del PRONIE MEP-FOD, en el que se considera importante la apertura al diálogo como un proceso de construcción de conocimiento. En general, es una estrategia fuerte dentro del campo de la definición de puntos de corte, con lo cual la amplia literatura existente podría ser oportuna para orientar los procesos que se hicieran desde el contexto de evaluación de la FOD.

- **Estrategia Bookmark: Pertinencia alta.**

Se considera pertinente para el contexto evaluativo de interés ya que coincide en cuanto al ámbito en el cual se utiliza y al tipo de ítems empleados. Otro aspecto importante es que permite determinar más de un punto de corte dentro de la prueba de evaluación, es decir, ofrecer una valoración por varios niveles de competencia, lo cual es coherente con los objetivos que se persiguen desde el modelo de evaluación de habilidades de los estudiantes del PRONIE MEP-FOD.

Se considera uno de los métodos que implica mayor rigor a nivel metodológico en comparación con las debilidades que se critican a otros métodos, lo cual se refleja en los controles y conocimientos estadísticos que se requieren para llevar a cabo su aplicación. En este sentido, el modelo es compatible con los modelos de evaluación, construcción y calibración de ítems que se han estado explorando e implementando en los últimos años en la Unidad de Evaluación de la FOD, como es el caso de los modelos de Teoría de Respuesta al Ítem. Partiendo de estas valoraciones esta estrategia es la que parece caracterizarse como la más pertinente según los objetivos evaluativos de la FOD.

C. Propuesta de implementación en el contexto evaluativo de la FOD

El análisis de pertinencia de las diferentes estrategias identificadas, arrojó información útil para elegir la estrategia **Bookmark** como la más oportuna para implementarse en el contexto de las pruebas de evaluación de la FOD. Cabe recordar que esta estrategia resultó ser la mejor puntuada en el análisis comparativo realizado por Aranguren y Hoszowski (2017), en el que se analizaron varios criterios de calidad y rigurosidad, lo cual da mayor solidez la selección de esta estrategia.

De acuerdo con Cizek y Bunch (2007) la estrategia *Bookmark* se considera como la sucesora de las técnicas de *item mapping* utilizadas por el NAEP y, en términos generales, consiste en la asignación de marcas en un Folleto ordenado de ítems (*Ordered Item Booklet -OIB*), que el juez realiza a partir de la revisión de un conjunto de ítems organizados previamente por orden de dificultad, según los parámetros de la Teoría de Respuesta al Ítem (*b value*). Como se observa en la **Figura 26**, en el OIB se coloca un ítem por página con su respectiva información (enunciado y opciones de respuesta) y cuya numeración corresponde al orden de aparición en la prueba. Además, se recomienda incluir la regla de probabilidad de acierto definida para la asignación del marcador.

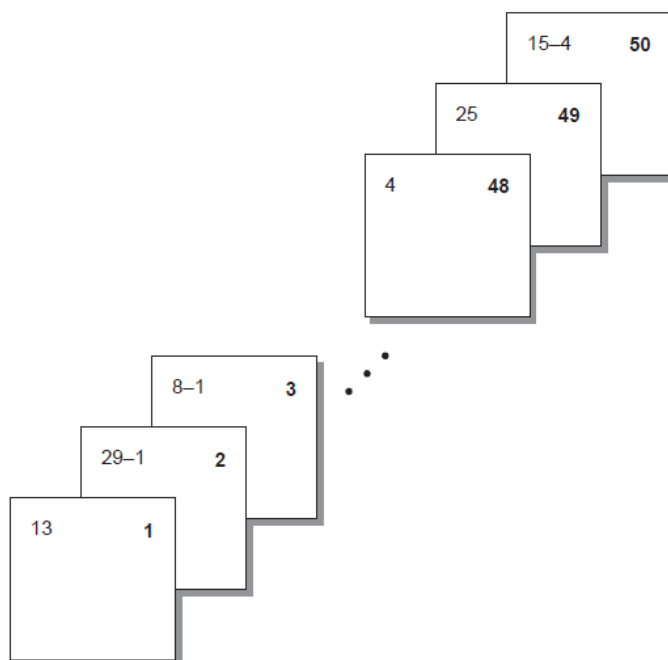


Figura 26. Ejemplo OIB en la estrategia Bookmark. (Fuente: Cizek y Bunch, 2007, p. 161).

En concordancia con los objetivos del TFM, en este apartado se derivan algunos criterios oportunos para implementar esta estrategia en el contexto de evaluación de competencias digitales del PRONIE MEP-FOD. Como se plantea en la **Figura 27**, la propuesta de implementación para dicho contexto se desarrolla en torno a cuatro fases y, posteriormente, se detallan los aspectos clave a considerar dentro de cada una de ellas.

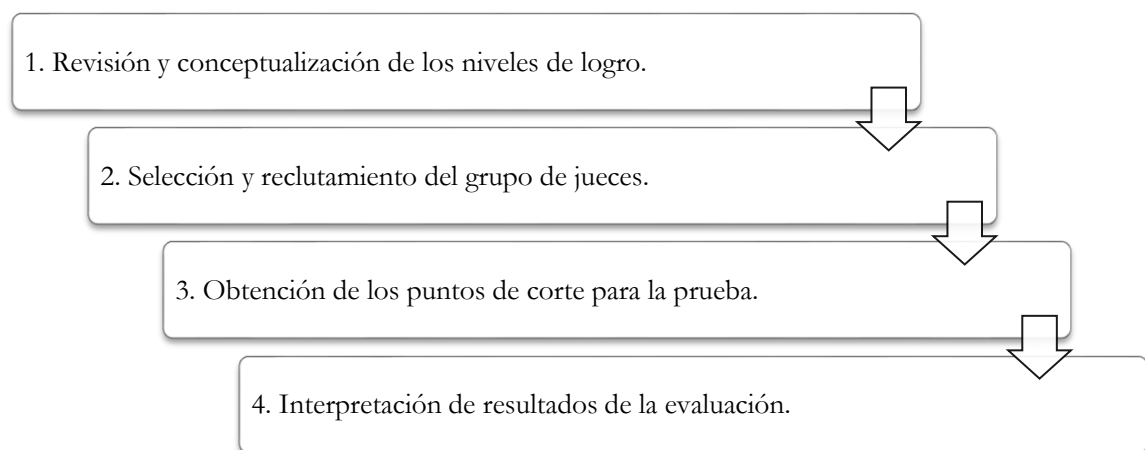


Figura 27. Fases de implementación de la estrategia *Bookmark* para definir los puntos de corte en la prueba de evaluación de la FOD. (Fuente: Elaboración propia a partir del conjunto de datos).

Las fases propuestas parten del supuesto de que la prueba de evaluación fue aplicada a los estudiantes, es decir, que ya se cuenta con los datos empíricos. Esto debido a que este aspecto es la base para la implementación de la estrategia *Bookmark* en tanto que se requiere de un análisis estadístico previo de los resultados de los evaluados para generar el ordenamiento de los ítems según la dificultad y organizar el OIB.

Fase 1. Revisión y conceptualización de los niveles de logro

Los procesos de evaluación de resultados de aprendizaje en estudiantes del PRONIE MEP-FOD han estado habitualmente orientados a aproximar tres niveles de logro, a partir de una distribución de los resultados de los y las estudiantes evaluadas por quintiles (Véase detalle en la **Figura 28**).

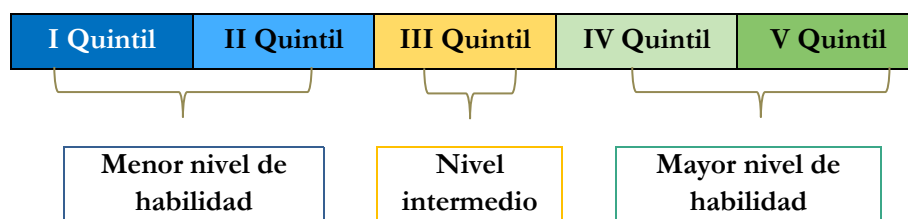


Figura 28. Categorización de los niveles de habilidad utilizados en las pruebas de la FOD, según quintiles (Fuente: FOD, 2017a, p.31).

Según se explica, esta distribución implica que (FOD, 2017b, pp. 31-32):

- ✓ Los estudiantes ubicados en el *I y II Quintil* corresponden a aquellos que tienen un *nivel inicial* de habilidad. Esto quiere decir que cuentan con un 50% de probabilidad de contestar correctamente los ítems de dificultad baja.
- ✓ Los estudiantes ubicados en el *III Quintil* tienen un *nivel intermedio* de habilidad. Es decir, que cuentan con un 50% de probabilidad de contestar correctamente los ítems de dificultad media y una alta probabilidad de acierto en los ítems de dificultad baja.
- ✓ Los estudiantes ubicados en el *IV y V Quintil* tienen un *nivel avanzado* de habilidad, lo cual implica que tienen un 50% de probabilidad de contestar correctamente los ítems más difíciles y una alta probabilidad de acierto en los ítems de los demás niveles.

Para la implementación de la estrategia *Bookmark* en este contexto, se recomienda trabajar con esos tres niveles asociados a los quintiles, sin embargo, se requieren algunas modificaciones. El primer paso necesario revisar y definir las etiquetas que definirían los niveles de competencia (Aranguren y Hoszowski, 2017). Estas deben ser simples y útiles para diferenciar entre las categorías de desempeño de manera neutral, clara y precisa.

En la literatura se sugieren algunos modelos que podrían ser útiles para el contexto de la FOD, por ejemplo, por nivel de desempeño [*Desempeño bajo, Desempeño medio y Desempeño alto*], [*Básico, Competente y Avanzado*], por nivel de habilidad [*Elemental, Satisfactoria, Sobresaliente*], por suficiencia [*Lejos del suficiente, Suficiente, Avanzado*] y por logro de objetivos [*No alcanza los objetivos, Alcanza los objetivos, Alcanza los objetivos satisfactoriamente*] (Aranguren y Hoszowski, 2017; Cizek y Bunch, 2007).

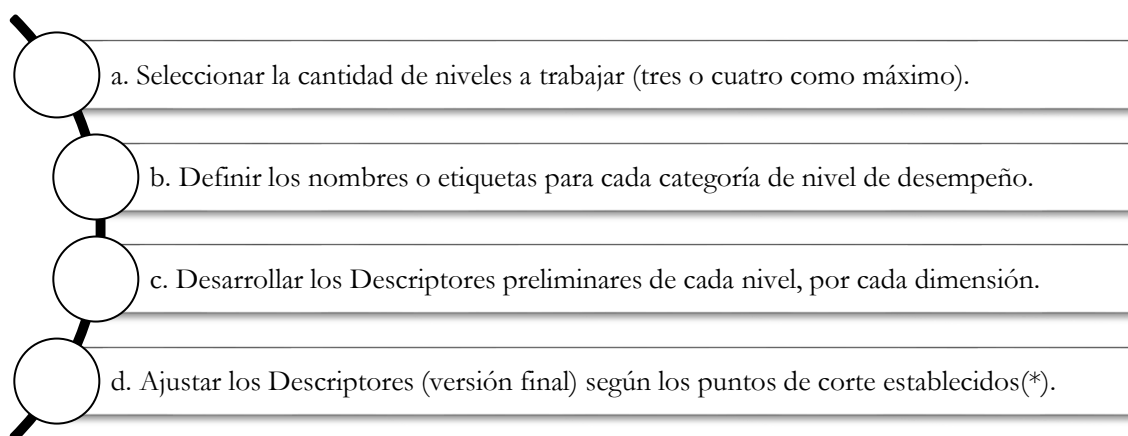
Una vez definidas las etiquetas, el próximo paso sería necesario abandonar la descripción actual de los niveles de la prueba que están enfocados en la probabilidad de acertar los ítems y trabajar en el desarrollo de unos *Descriptor del nivel de habilidad* que caracterizarían

de manera preliminar lo que se espera de los y las estudiantes en cada nivel de logro. Estos descriptores deberían hacerse para ambas dimensiones a estudiar, ya que estarían directamente relacionados con los contenidos particulares de cada una y con las implicaciones de cada Estándar de Desempeño vinculado.

Los descriptores del nivel de habilidad son importantes dentro del proceso de obtención de puntos de corte no sólo porque son útiles para orientar el trabajo que realizan los jueces, sino porque finalmente “el significado” de los resultados obtenidos, es decir, que son la base para la interpretación de los resultados de la prueba. Esto hace que su conceptualización sea un paso al que se le debe hacer una adecuada inversión de tiempo por parte de los encargados de la prueba, que en este caso, corresponde al equipo de trabajo de la Unidad de Evaluación de la FOD.

Se considera que dicha Unidad tiene trabajo adelantado alrededor de este proceso de generación de los descriptores, en tanto que ha desarrollado la prueba basada en unos indicadores de desempeño específicos para cada estándar, que además, están contruidos en progresión. De esta manera, la recomendación es tomar dichos indicadores como punto de partida para desarrollar los descriptores provisionales de cada nivel de logro. Sin embargo, se debe considerar que la versión definitiva de estos descriptores se desarrolla luego de haber establecido los puntos de corte definitivos que separan cada categoría.

De esta manera, en la **Figura 29** se sintetizan los pasos asociados a esta fase.



(*): Paso posterior a la finalización de la fase de establecimiento de puntos de corte.

Figura 29. Síntesis de pasos asociados a la Fase de revisión y conceptualización de los niveles de logro. (Fuente: Elaboración propia a partir del conjunto de datos).

Fase 2. Selección y capacitación del grupo de jueces

Aunque no se encuentra en la literatura un número concreto como sugerencia para determinar la cantidad de jueces que deberían conformar un grupo de trabajo dentro de la implementación de una estrategia de definición de puntos de corte (Cizek y Bunch, 2007), se sugiere que sea un número representativo en términos de ruralidad y tipo de centro educativo. Asimismo, se recomienda contar una muestra de jueces que involucre ambos sexos, varios rangos de edad y diferentes años de experiencia laboral (Aranguren y Hoszowski, 2017).

La selección de los jueces debe partir entonces de dichas características, pero a la vez de la necesidad de que conozcan a profundidad los contenidos a evaluar. Es por ello que en el caso de la evaluación de estudiantes del PRONIE MEP-FOD, se considera necesario contar con un grupo de jueces que esté integrado por Docentes de Informática Educativa y de Asesores Nacionales de Informática Educativa.

Tomando uno de los ejemplos de conformación de grupos de jueces sugeridos por la experiencia de Aranguren y Hoszowski (2017) se considera viable para la FOD contar con un panel de entre 18 y 24 jueces, para facilitar las discusiones en subgrupos de al menos cuatro participantes. Para ello, se recomienda contar con un grupo inicial de al menos 30 personas, como un respaldo ante cualquier dificultad durante las sesiones.

Por otra parte, uno de los procesos más relevantes dentro de esta fase es la capacitación que se realiza al grupo de jueces, ya que el éxito de la implementación de la estrategia depende en gran medida de la adecuada comprensión de éstos sobre la tarea (Cizek y Bunch, 2007). Al respecto, Hambleton (2001), destaca algunos pasos necesarios para llevar cabo un proceso de capacitación de esta naturaleza, con lo cual se destacan los principales aspectos y se amplía en función de la implementación de la estrategia *Bookmark* en el caso de la FOD.

- 1. Explicar los pasos a seguir para establecer estándares.** Se recomienda hacer una contextualización a los jueces sobre la importancia de contar con unos niveles de logro para interpretar los resultados de las evaluaciones. Es importante brindar información sobre lo que se espera de su parte durante las sesiones, sobre las etiquetas y sobre los descriptores preliminares generados por cada nivel y para cada dimensión de los Estándares de Desempeño. Adicionalmente, se recomienda llevar a cabo una

comprobación de la comprensión que tuvieron los jueces sobre la contextualización inicial para asegurar que visualizan la importancia y el alcance de la tarea.

2. Presentar los documentos en donde deberán establecer sus puntajes de corte.

En el caso de la implementación de la estrategia *Bookmark*, se requiere de la elaboración previa del OIB, de los formularios de respuesta y los marcadores (que pueden ser autoadhesivos de tipo *post-it*). De esta manera, durante el proceso de capacitación se dará la oportunidad a los jueces de familiarizarse con estos insumos.

Como se ha detallado, en el OIB se muestra el conjunto de ítems ordenados según el nivel de dificultad, mientras que, el formulario de respuesta es un material en el que se sintetizan los juicios emitidos por cada juez y que incluye un espacio de información sobre la prueba y un espacio de respuesta por cada ronda de emisión de juicios, que dentro de la estrategia *Bookmark* pueden ser hasta tres (Véase ejemplo en el **Anexo 8**).

3. Practicar la determinación de un punto de corte. Otro aspecto clave a trabajar es la conceptualización del sujeto con mínima habilidad o *sustentante límite*. Este concepto es sumamente importante de ser abordado con los jueces para garantizar que comprenden sobre qué base deben formular sus valoraciones. Además, se debe trabajar en la clarificación de la regla para ubicar el marcador, que en la estrategia *Bookmark* se recomienda utilizar la de $2/3$ (correspondiente al 0,67 de probabilidad de acierto en un modelo de 3PL o de 0,50 en uno de 2PL). La consigna para el juez es: *Indicar en qué punto (ítem) consideran que la probabilidad de que un sujeto con un mínimo de habilidad responda el ítem de manera correcta sería menor a 0,67, es decir, que caería por debajo de la regla establecida.*

Partiendo de la clarificación de estos conceptos y de la comprobación de su comprensión por parte de los jueces es oportuno llevar a cabo un proceso de práctica en cuanto a la asignación de un punto de corte. De esta manera, los jueces tienen la oportunidad de poner en común su comprensión del proceso y de los materiales, así como de resolver en conjunto algunas de las dudas que hayan surgido en relación con la tarea requerida por la estrategia.

4. Repasar el conjunto de ítems sobre el cual se aplicará el procedimiento. Dentro del proceso de capacitación es importante que los jueces tengan un espacio para familiarizarse con los tipos de ítems correspondientes a las dos dimensiones evaluadas.

En este caso, se puede reservar un espacio para que los jueces completen un ejemplar de la prueba de evaluación empleada.

Los aspectos planteados ayudan a visualizar la importancia que el proceso de capacitación de los jueces tiene dentro de la implementación de la estrategia *Bookmark*. Como consideraciones adicionales, destaca la recomendación de hacer las sesiones de manera presencial por lo que a nivel de la FOD, se tendrían que tomar las previsiones logísticas que esto implica para el trabajo con los jueces seleccionados. Asimismo, en la **Figura 30** se muestra una síntesis de los pasos que deberían contemplarse dentro de esta segunda fase de implementación de la estrategia seleccionada.

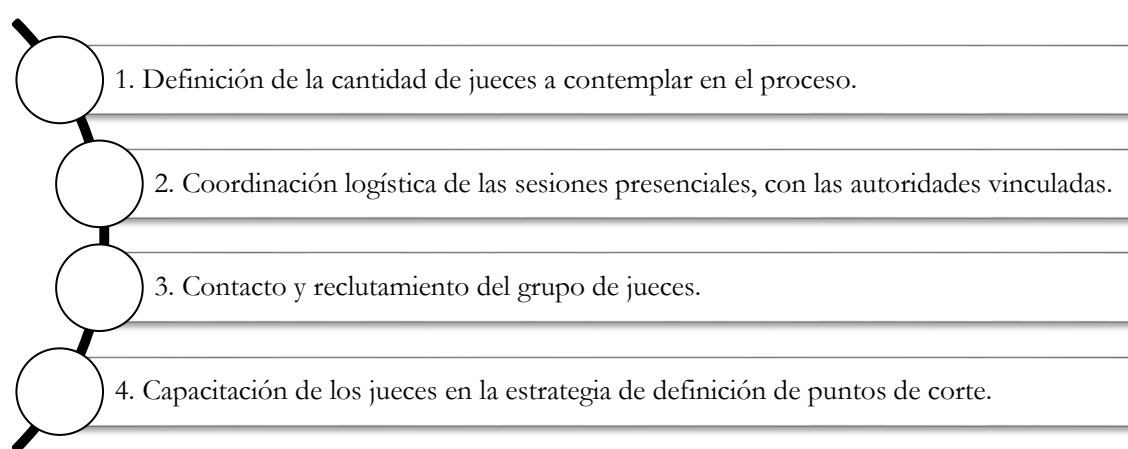


Figura 30. Síntesis de pasos asociados a la Fase de selección y capacitación de los jueces (Fuente: Elaboración propia a partir del conjunto de datos).

Fase 3. Obtención de los puntos de corte para la prueba

Esta fase corresponde al desarrollo de los pasos específicos para obtener los puntos de corte que implica la estrategia *Bookmark*. Antes de entrar en el detalle de dichos pasos, es importante mencionar que este procedimiento requiere llevarse a cabo de manera presencial, con lo cual se recomienda pensar en una jornada de dos o tres días en los que se lleve a cabo la capacitación y, posteriormente, el trabajo concreto de esta fase (Véase propuesta de agenda de trabajo en el **Anexo 9**).

El procedimiento detallado para obtener los puntos de corte mediante la estrategia *Bookmark* se divide en tres rondas de valoración por parte de los jueces, las cuales se detallan en la **Tabla 16**.

Tabla 16. Rondas de valoración a realizar durante la implementación de la estrategia *Bookmark*

Valoración	Procedimiento	Resultados
Ronda 1	1. A cada juez se le entrega el material requerido (OIB, formulario de respuesta y marcadores).	✓ Primeros marcadores ubicados.
	2. En subgrupos de tres o cinco discuten sobre lo que cada ítem aproxima y lo que implica en cuanto a conocimiento o habilidad para la población estudiantil.	✓ Primeros puntos de corte identificados.
	3. De manera individual, el juez coloca el marcador en el punto que considera que la regla sobre la probabilidad de acierto deja de cumplirse.	
	4. Los jueces completan el formulario de respuestas.	
	5. Se calculan los puntos de corte transformando las marcas establecidas por los jueces en puntajes escala correspondiente con el nivel de habilidad.	
	6. Se repite el procedimiento para cada dimensión.	
Ronda 2	7. Se muestran los resultados de la primera ronda de juicios a los jueces para que puedan visualizar y discutir las discrepancias y similitudes en conjunto (realimentación con Información normativa del grupo).	✓ Revisión de la colocación de los marcadores.
	8. Además del material inicial, se entrega un resumen de los resultados de la primera ronda de juicios.	✓ Versión preliminar de los puntos de corte.
	9. Se conforman nuevos subgrupos para la discusión y se repiten los pasos 3, 4, 5 y 6 (también pueden dejarse los mismos grupos si se considera oportuno).	
Ronda 3	10. Se muestran los resultados de la segunda ronda de juicios (Información normativa) y la cantidad de estudiantes que se asignarían a cada nivel de logro según los puntos de corte preliminares (Información de impacto).	✓ Revisión final de los marcadores.
	11. Se genera discusión grupal y se repiten los pasos 4, 5 y 6.	✓ Versión final de los puntos de corte para cada dimensión.
	12. La media de los puntos de corte asignados como VF por cada juez, corresponde al punto de corte final por cada nivel y para cada dimensión.	

(Fuente: Elaboración propia a partir de las descripciones del procedimiento elaboradas por Aranguren y Hoszowski, 2017 y Cizek y Bunch, 2007).

En términos generales, la estrategia promueve las discusiones entre los jueces, pero parte de la asignación individual de los marcadores. Así, el establecimiento de los puntos de corte finales se realizan en función de la sumatoria y promedio de los valores de habilidad (*theta*)

correspondientes a la ubicación de los marcadores que hicieron los jueces. Cabe mencionar que, de acuerdo con Hambleton (2001), los puntos de corte finales pueden ser levemente modificados para unificar criterios, en aquellos casos en los que se haya contado con varios grupos de jueces.

Una vez obtenidos los puntos de corte definitivos, se debe trabajar en la elaboración final de los Descriptores de desempeño de cada nivel. En esta tarea puede contarse con la participación de los jueces, además del equipo técnico encargado del proceso. Una recomendación útil para la FOD sería discutir en conjunto con jueces los aspectos a contemplar en cada Descriptor de nivel de logro según los puntos de corte establecidos y, posteriormente, enviar las redacciones finales de los descriptores por correo electrónico para que los jueces emitan sus valoraciones sobre éstos.

La sesión de trabajo con los jueces debe terminarse con una valoración por parte de éstos para efectos de garantizar algunos controles en cuanto a la confiabilidad y validez del procedimiento. En el **Anexo 10** se muestra un ejemplo de escala de valoración propuesto por Cizek et al (citados en Aranguren y Hoszowski, 2017), en el que se contemplan aspectos como la claridad de las instrucciones recibidas, la comprensión de las tareas requeridas y la utilidad de las discusiones, entre otros aspectos.

Partiendo del detalle del procedimiento descrito con anterioridad, en la **Figura 31** se muestra la síntesis de pasos correspondientes a esta tercera fase.

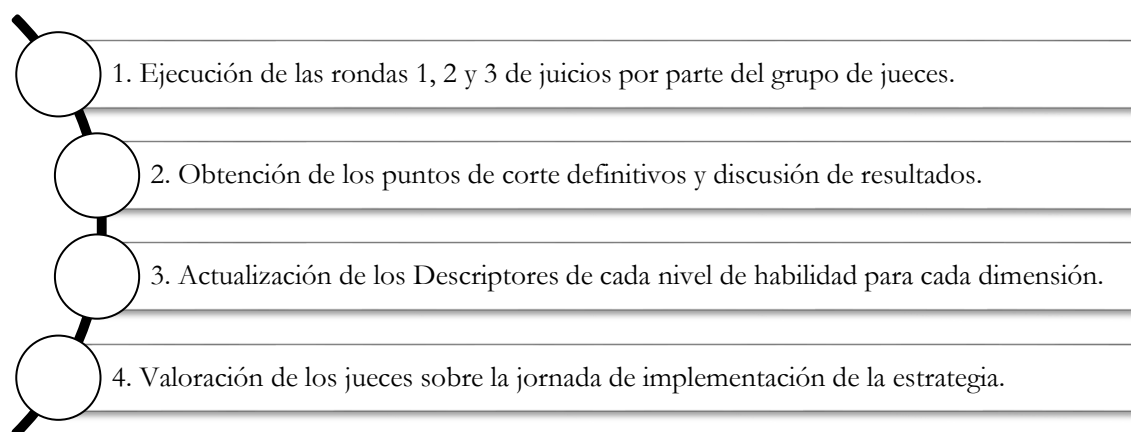


Figura 31. Síntesis de pasos asociados a la Fase de obtención de los puntos de corte. (Fuente: Elaboración propia a partir del conjunto de datos).

Fase 4. Interpretación de resultados de la evaluación

La última fase corresponde a la interpretación de los resultados obtenidos por los y las estudiantes en la prueba, esta vez, a partir de los niveles de logro obtenidos a partir del establecimiento de los puntos de corte. Al respecto, será necesario describir los resultados obtenidos para cada una de las dimensiones evaluadas que, en este caso, corresponden a las de Productividad y Resolución de problemas e investigación. Concretamente, se esperaría identificar la distribución de los evaluados en cada nivel de logro de cada dimensión y, de esta manera, generar resultados en torno a los perfiles de los y las evaluadas.

Según se detalla a FOD (2018), posteriormente estos niveles de logro se tomarían como variable dependiente para llevar a cabo un análisis de regresión logística que permita aproximar la asociación de otras variables con dichos niveles de logro. Algunas de las variables que se contemplarían se relacionan con la experiencia de los y las estudiantes como beneficiarios del PRONIE MEP-FOD, algunas de sus características personales y de su centro educativo, entre otras. En la **Figura 32** se muestra el detalle de las variables a contemplar.

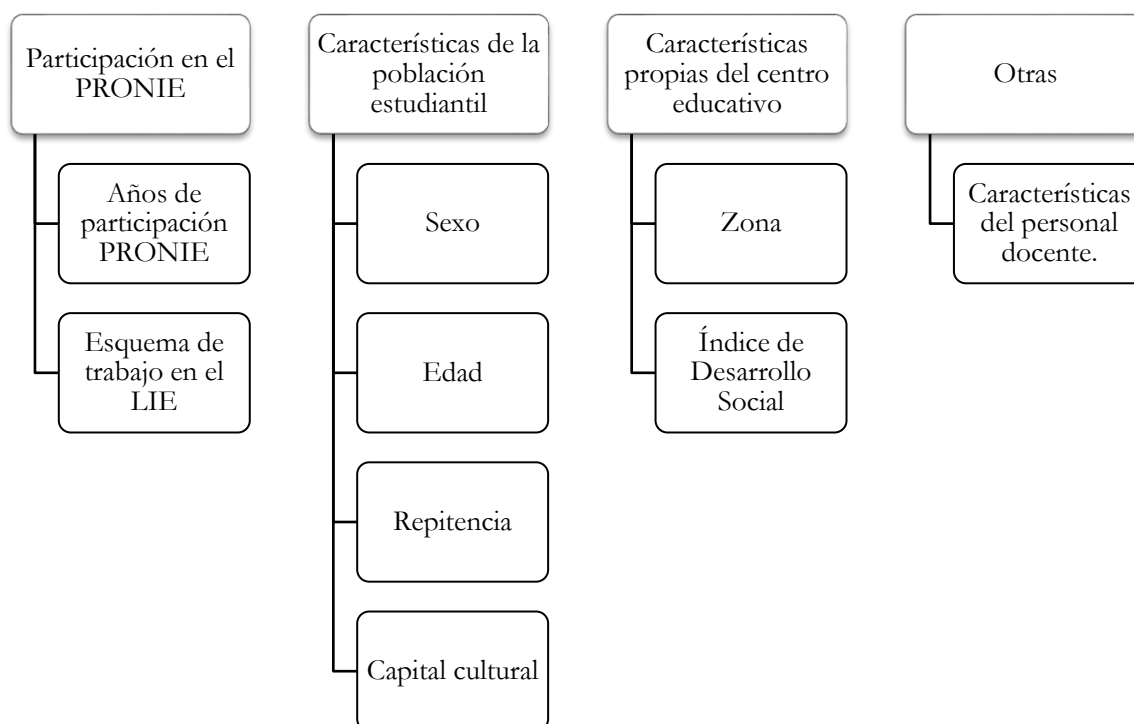


Figura 32. Variables independientes a considerar en la profundización de factores asociados a los niveles de logro de los estudiantes en la prueba. (Fuente: FOD, 2018).

Finalmente, cabe valorar la utilidad de realizar un pilotaje del proceso de implementación de la estrategia utilizando, por ejemplo, una de las dimensiones evaluadas en el año 2017. De esta manera, se tendría la oportunidad de que el equipo de evaluación de la FOD se familiarice con el método, a la vez que se podrían tomar algunas otras previsiones oportunas para la ejecución formal del proceso con los datos que se obtengan del proceso de aplicación de la prueba de evaluación durante el año 2018.

V. Conclusiones

En el siguiente apartado se destacan las principales conclusiones de esta investigación. Asimismo, se plantean los principales alcances y limitaciones identificados y se puntualizan algunas recomendaciones para futuras investigaciones vinculadas con el tema o la metodología empleada en este estudio.

A. Sobre los resultados del estudio

La configuración intelectual del campo de conocimiento de definición de puntos de corte en pruebas de evaluación referidas a criterios es un tema desarrollado desde hace varias décadas (finales de los 70) pero que ha tenido su mayor crecimiento en cuanto a volumen de producción científica a partir del año 2000. Asimismo, al valorar la evolución temporal del campo de conocimiento se evidenció una progresión desde aspectos vinculados a la metodología en los primeros años, hacia la especificación en campos de evaluación de conocimiento concretos como el de educación y el de la salud.

El análisis de tendencias realizado mediante la estrategia de *Science Mapping* ubica el conjunto de datos dentro del ámbito de la educación. Asimismo, dentro de los principales hallazgos, se destaca Estados Unidos como el país más productivo dentro del campo de conocimiento, lo cual es coherente con las tendencias actuales en cuanto los países que son potencias en temas de investigación en educación. Sin embargo, llama la atención la ausencia de países latinoamericanos dentro del conjunto de referentes clave, lo cual sugiere que el campo podría ser aun incipiente dentro de dicho contexto (al menos en cuanto a producción científica).

Otro hallazgo relevante es que la semántica del conjunto de datos es coherente con la selección de descriptores utilizados en la búsqueda de información en la WOS. Esto se refuerza con el contenido de los tres subgrupos temáticos en torno a los cuales interactuaron las publicaciones, correspondientes a temas educativos y metodológicos vinculados a las evaluaciones de desempeño de estudiantes.

En términos generales, se evidencia que el análisis de la configuración intelectual del campo de conocimiento de la definición de puntos de corte en pruebas de evaluación referidas a criterios permitió realizar un mapeo de sus principales características, conformación y evolución temporal. Al mismo tiempo, permitió identificar las principales tendencias en

cuanto a citación y países más productivos con lo que se logra obtener una importante descripción de la configuración intelectual asociada a dicho campo de conocimiento.

Además de constituir un importante hallazgo en sí misma, esta exploración del campo de conocimiento funcionó como base para llevar a cabo un análisis en profundidad de las estrategias de definición de puntos de corte identificadas en la literatura. Dicho análisis mostró que la mayoría de estrategias se utiliza efectivamente en el ámbito educativo y que requiere contar con datos empíricos para implementarse, además de trabajar principalmente con ítems de selección única. También se identificó que la mayoría es útil para definir puntos de corte dicotómicos y, en menor medida, se identificaron estrategias útiles para clasificar a los sujetos en varios niveles de logro.

Las características anteriormente discutidas, así como otras de las contempladas en el análisis de pertinencia realizado en función de los requerimientos de la prueba de evaluación de la FOD, permitieron descartar 10 de las 17 estrategias de definición de puntos de corte identificadas. Entre los principales criterios que definieron la categorización de estrategias como aquellas con Pertinencia Baja destacaron debilidades a nivel metodológico, la determinación de puntos de corte dicotómicos y el requerir que el juez fuera directamente el docente del evaluado. Esta última consideración se tomó en cuenta al valorar que en pruebas de evaluación a gran escala, como es el caso de las pruebas de evaluación de la FOD, resulta poco viable reclutar a todos los docentes del total de participantes en los estudios a nivel país.

Por su parte, otras estrategias mostraron concordancia en algunos de los requerimientos asociados a la prueba de evaluación de la FOD, pero tenían alguna característica que las hacía menos pertinentes. De esta manera, se clasificaron cinco de las 17 estrategias como con Pertinencia Media. Asimismo, el análisis permitió identificar las dos estrategias restantes como las que tenían un nivel de Pertinencia Alta, es decir, que cumplían con la mayoría de los requerimientos identificados como clave dentro del contexto de evaluación de estudiantes beneficiarios del PRONIE MEP-FOD.

Partiendo de los criterios contemplados en la categoría de Pertinencia Alta como el permitir establecer varios niveles en la prueba, el admitir ítems de selección única, el no requerir que los jueces fueran directamente los docentes de cada uno de los evaluados y el mostrar solidez metodológica, se eligió la estrategia *Bookmark* como más adecuada. De esta manera, se genera un tercer resultado relevante dentro del TFM, el cual corresponde a la propuesta

de implementación de dicha estrategia según los requerimientos del contexto de evaluación de la FOD.

De manera global, se considera que el trabajo que se ha estado realizando desde la Unidad de Evaluación de la FOD en torno a la evaluación de estudiantes significa un avance importante para la implementación de la estrategia *Bookmark*. Sin embargo, se requiere ajustar algunos aspectos a las demandas de la estrategia para lo cual se realiza una propuesta de implementación en cuatro fases, dentro de las que se plantean consideraciones logísticas, procedimentales y de análisis que se requieren para llevar a cabo la obtención de los puntos de corte en la prueba de evaluación de estudiantes del PRONIE MEP-FOD.

Finalmente, un aspecto que se destaca es la recomendación al equipo FOD de brindar al equipo la oportunidad de llevar a cabo un pilotaje de dicho procedimiento, antes de implementar la de manera oficial con el grupo de jueces. Esto último con el objetivo de identificar aspectos que podrían favorecer la implementación, así como controlar aquellos que podrían limitarla y, en general, favorecer la familiarización del equipo con las técnicas de análisis vinculadas con esta estrategia.

B. Alcances y limitaciones

Los principales aportes del presente estudio se orientan a contribuir con la mejora de los procesos de evaluación desarrollados en el contexto de la Fundación Omar Dengo, en Costa Rica, así como a documentar una experiencia de exploración de un campo de conocimiento a través de una estrategia de revisión creativa de la literatura como es el *Science Mapping*. Así, la contribución no sólo se enfoca en aportar conocimientos teóricos sobre el campo de conocimiento, sino que se buscó ampliar mediante un análisis en profundidad de las estrategias identificadas con el fin de orientar la toma de decisiones de la Unidad de Evaluación en torno al proceso de selección de una estrategia oportuna, según los requerimientos de dicho contexto.

El desarrollo del TFM contribuye, además, con una serie de consideraciones metodológicas oportunas para conducir un proceso de implementación de una estrategia de definición de puntos de corte en un contexto educativo de evaluación de competencias digitales. De esta manera, se pretende que los conocimientos derivados sean útiles para optimizar los procesos de implementación, calendarizados para el segundo semestre del año 2018 en el

marco de la evaluación anual de los estudiantes beneficiarios de los Laboratorios de Informática Educativa del PRONIE MEP-FOD.

Por su parte, la utilización de la estrategia de *Science Mapping* como hilo orientador del TFM para realizar la exploración del campo de conocimiento permitió llevar a cabo un análisis de la configuración intelectual del mismo que no sería posible de alcanzar mediante otras estrategias de análisis bibliográfico. De esta manera, el estudio ofrece un ejemplo del alcance de este tipo de estrategias de revisión creativa de la literatura como vía efectiva y pertinente para la exploración de una temática concreta en medio de las grandes cantidades de información de las que se dispone en la actualidad.

El potencial de aprovechamiento de las bases de datos científicas y de las diferentes herramientas de software que se han desarrollado en los últimos años para apoyar su uso, se refleja en el diseño metodológico planteado en este estudio. Así, el presente TFM contribuye con evidencia empírica sobre cómo los avances tecnológicos pueden ser utilizados de manera oportuna en el campo de la investigación aplicada a la educación. En esta misma línea, se considera que el establecimiento de criterios de calidad y la definición de parámetros concretos para llevar a cabo el análisis son oportunos para conducir líneas de investigación similares, es decir, que estén enmarcadas dentro de procesos de revisión creativa de la literatura.

Por otra parte, se destacan algunas de las limitaciones propias del estudio. Una de ellas es la ausencia de referentes propios del contexto latinoamericano, ya que dentro del conjunto de datos no se encontró mayor representación de este contexto. Aunque se cuenta con registro de algunas experiencias, los resultados evidencian que parece ser que existe una ruptura entre lo que se hace y lo que se publica dentro de la comunidad científica. Esta limitación se encuentra también en el contexto de la FOD, desde el cual se han estado elaborando importantes procesos de investigación que incluso podrían estar siendo pioneros en la región, pero que no se visualizan de manera tan fuerte a nivel de comunidad científica, sino que el acceso a la información únicamente es posible si se cuenta con acceso a la documentación institucional interna, como es este el caso.

Ante este escenario cabe destacar los esfuerzos que se han estado realizado recientemente para aumentar el volumen de publicación científica vinculada con los procesos de investigación en estos contextos. La relevancia de esto se orienta a la posibilidad de que otros estudios similares al presente TFM puedan contar con mayor cantidad de

experiencias documentadas en el contexto latinoamericano, de manera que se puedan enriquecer las propuestas y análisis generados.

Finalmente, el análisis de la configuración intelectual del campo de conocimiento muestra de manera amplia el tipo de análisis que se podría realizar. Sin embargo, se considera que el estudio se podría mejorar incluyendo algunos otros parámetros de análisis más complejos que actualmente están disponibles dentro de este tipo de estudios bibliométricos en los que se utilizan estrategias como las de *Science Mapping*. Lo anterior, da pie a la formulación de algunas líneas de investigación futura posibles derivadas del presente estudio.

C. Recomendaciones y líneas de investigación futura

- ✓ Sobre la aplicabilidad de los resultados de esta investigación, se recomienda utilizarlos para optimizar los procesos de evaluación de estudiantes beneficiarios de las propuestas de Informática Educativa del PRONIE MEP-FOD. Sin embargo, para la implementación en otros contextos de evaluación sería necesario llevar a cabo un nuevo análisis de pertinencia que permitiera seleccionar la estrategia más adecuada en función de los criterios y los requerimientos propios del contexto particular en el que se quiera implementar.
- ✓ Se recomienda continuar con esta línea de investigación en el contexto educativo costarricense en el que se desarrollan las pruebas de evaluación sobre las cuales se fundamenta el estudio, es decir, desde la FOD. De esta manera se amplía la posibilidad de dar solidez a los hallazgos del presente estudio a partir, por ejemplo, de la comprobación del funcionamiento de la estrategia seleccionada y de la identificación de algunas otras condiciones necesarias para su adecuada implementación.
- ✓ El presente estudio sirve como base para la implementación de la estrategia de revisión creativa de la literatura conocida como *Science Mapping*, por lo que se recomienda que futuras investigaciones puedan ampliar las posibilidades de profundización en la caracterización que se realiza de la configuración intelectual de un campo de conocimiento, a partir, por ejemplo, de la exploración y trabajo con parámetros más complejos que actualmente están disponibles dentro del repertorio de los análisis bibliométricos.

- ✓ En esa misma línea, se considera interesante realizar una comparación de los alcances y limitaciones de las principales herramientas de software que actualmente se encuentran disponibles para llevar a cabo análisis bibliométricos de este tipo. De esta manera, se podría generar mayor orientación en cuanto a los criterios o pautas útiles a tomar en cuenta al realizar un análisis de este tipo.

- ✓ Finalmente, dentro del contexto específico del Máster de Investigación aplicada a la Educación, el presente TFM abre la puerta a la exploración de otras formas y estrategias metodológicas oportunas que evidencien el aprovechamiento posible del potencial de las herramientas tecnológicas disponibles en la actualidad, las cuales están al servicio de la investigación educativa pero no son ampliamente exploradas ni documentadas.

VI. Referencias

- Ala-Mutka, K. (2011). *Mapping digital competence: Towards a conceptual understanding*. Luxemburgo: JRC-IPTS European Commission. Recuperado de <http://ipts.jrc.ec.europa.eu/publications/pub.cfm?id=4699>
- American Educational Research Association [AERA]. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
- Angarita, L. (2014). Estudio bibliométrico sobre uso de métodos y técnicas cualitativas en investigación publicada en bases de datos de uso común entre el 2011-2013. *Revista Iberoamericana de Psicología: Ciencia y Tecnología*, 7(2), 67-76.
- Angoff, W. (1971). *Scales, norms, and equivalent scores*. In R. L. Thorndike (Ed.), *Educational measurement* (pp. 508–600). Washington, DC: American Council on Education.
- Aranguren, M. y Hoszowski, A. (2017). *Aprender 2016. Serie de documentos técnicos 5: Bookmark, Establecimiento de puntos de corte*. Ministerio de Educación y Deportes: Buenos Aires. Disponible en http://www.educacion.gob.ar/data_storage/file/documents/manual-bookmark-595bd361cf4e7.pdf
- Aria, M. & Cuccurullo, C. (2017) Bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics*, 11(4), 959-975.
- Barbosa, J., Barbosa J. y Rodríguez, M. (2013). Revisión y análisis documental para estado del arte: Una propuesta metodológica desde el contexto de la sistematización de experiencias educativas. *Investigación Bibliotecológica*, 27(61), 83-105.
- Barquero, K. (07 de agosto, 2017). Costa Rica invierte en educación más que cualquier país de OCDE. *La República*. Recuperado de <https://www.larepublica.net/noticia/costa-rica-invierte-en-educacion-mas-que-cualquier-pais-de-ocde>
- Barrios, M. y Cosculluela, A. (2013). Fiabilidad. En J. Meneses, M. Barrios, A. Bonillo, A. Cosculluela, L. Lozano, J. Turbany y S. Valero (Eds.). *Psicometría* (pp. 75-141). Barcelona, España: OUC.

- Berk, R. (1986). A consumer's guide to setting performance standards on criterion referenced tests. *Review of Educational Research*, 56, 137–172.
- Bernal, S., Martínez, M., Parra, A. y Jiménez, J. (2015). Investigación documental sobre calidad de la educación en instituciones educativas del contexto Iberoamericano. *Revista Entramados- Educación y Sociedad*, 2(2), 107- 124.
- Beuk, C. (1984). A method for reaching a compromise between absolute and relative standards in examinations. *Journal of Educational Measurement*, 21, 147–152.
- Buckendahl, C., Smith, R., Impara, J. & Plake, B. (2002). A comparison of Angoff and Bookmark Standard Setting Methods. *Journal of Educational Measurement*, 39(3), 253-263.
- Cetin, S. & Gelbal, S. (2013). A comparison of Bookmark and Angoff Standard Setting Methods. *Educational Sciences: Theory & Practice*, 13(4), 2169-2175.
doi:10.12738/estp.2013.4.1829
- Chen, Ch. (2017). Science Mapping: A Systematic Review of the Literature. *Journal of Data and Information Science*, 2(2), 1–40.
- Chen, Ch. (2018). *How to Use CiteSpace*. [e-book]: Lean Publishing.
- Cizek, G. & Bunch, M. (2007). *Standard setting: A guide to establish and evaluating performance standards on tests*. California: Sage Publications.
- Cizek, G., Bunch, M. & Koons, H. (2004). Setting performance standards: Contemporary methods. *Educational Measurement: Issues and Practice*, 23(4), 31-49. doi:10.1111/j.1745-3992.2004.tb00166.x
- Constitución Política de la República de Costa Rica (1949). *Artículo 78 del Título VIII: La Educación y la Cultura*. Recuperado de <http://www.tse.go.cr/pdf/normativa/constitucion.pdf>
- Cortés, J. (2008). Web of Science: Termómetro de la producción internacional de conocimiento: Ventajas y limitaciones. *Culyt*, 5(29), 5-15.
- ENLACES (2012). *Niveles de logro 2º medio SIMCE TIC 2011*. Ministerio de Educación, Centro de Educación y Tecnología: Santiago de Chile.

- Esquivel, J. (2001). El diseño de las pruebas para medir logro académico: ¿referencia a normas o a criterios? En P. Ravela (Ed.), R. Wolfe, G. Valverde y J. Esquivel. *Los próximos pasos: ¿Cómo avanzar en la evaluación de los aprendizajes en América Latina?* (pp. 20-29). Programa de Promoción de la Reforma Educativa en América Latina y el Caribe [PREAL].
- Ferrari, A. (2012). *Digital competence in practice: An analysis of frameworks*. JRC Technical Reports. Joint Research Center. European Commission. Recuperado de <http://ftp.jrc.es/EURdoc/JRC68116.pdf>.
- Ferrari, A. (2013). *DIGCOMP: A Framework for Developing and Understanding Digital Competence in Europe*. Sevilla: JRC-IPTS. Recuperado de <http://ipts.jrc.ec.europa.eu/publications/pub.cfm?id=6359>
- Fundación Omar Dengo [FOD]. (2009). *Estándares de desempeño de estudiantes en el aprendizaje con tecnologías digitales*. Fundación Omar Dengo: San José, Costa Rica. Disponible en: http://www.fod.ac.cr/estandares/docs/estandares_desempeno.pdf
- Fundación Omar Dengo [FOD]. (2014a). *Nivel de logro de los Estándares de desempeño en estudiantes egresados del Segundo ciclo del PRONIE MEP-FOD* (Documento interno). Unidad de Evaluación, Fundación Omar Dengo, San José, Costa Rica.
- Fundación Omar Dengo [FOD]. (2014b). *Antecedentes y principios metodológicos: Actualización de la propuesta educativa del PRONIE MEP-FOD* (Documento interno). San José, Costa Rica: FOD.
- Fundación Omar Dengo [FOD]. (2015a). *Diagnóstico del nivel de logro de los Estándares de Desempeño de estudiantes de Tercer Ciclo en el aprendizaje con tecnologías digitales*. (Documento interno). Unidad de Evaluación, Fundación Omar Dengo, San José, Costa Rica.
- Fundación Omar Dengo [FOD]. (2015b). *Evaluación de los estándares de desempeño de estudiantes en el aprendizaje con tecnologías digitales, PRONIE MEP-FOD: Marco de Referencia*. (Documento interno). Unidad de Evaluación, Fundación Omar Dengo, San José, Costa Rica.
- Fundación Omar Dengo [FOD]. (2017a). *Nivel de logro de los Estándares de Desempeño en estudiantes de sexto grado del PRONIE MEP-FOD 2016: Informe de resultados* (Documento interno). Unidad de Evaluación, Fundación Omar Dengo, San José, Costa Rica.

- Fundación Omar Dengo [FOD]. (2017b). *Informe semestral de monitoreo institucional*. (Documento interno). San José, Costa Rica: FOD.
- Fundación Omar Dengo [FOD]. (2018). *Aporte de la propuesta de Laboratorios de Informática Educativa en la promoción de habilidades de productividad y resolución de problemas e investigación: Diseño de investigación*. (Documento interno). San José, Costa Rica: FOD.
- García, F. (2015). *Investigación documental: Leer, pensar y hablar con respecto de un tema definido para escribir bien y con provecho*. México: Limusa.
- García, P., Abad, F., Olea, J. y Aguado, D. (2012). *Un nuevo método de standard setting basado en la TRI: aplicación a eCat-Listening*. Informe de investigación eCAT 12-01. Madrid, España: Universidad Autónoma de Madrid.
- García-Varcárcel, A. (2016). *Las competencias digitales en el ámbito educativo*. Monografías del Departamento de Didáctica, Organización y Métodos de Investigación: Universidad de Salamanca. Disponible en <https://gredos.usal.es/jspui/bitstream/10366/130340/1/Las%20competencias%20digitales%20en%20el%20ambito%20educativo.pdf>
- Garza, J. (11 de enero de 2016). Costa Rica, único libre de analfabetismo en Centroamérica. *La República*. Recuperado de https://www.larepublica.net/noticia/costa_rica_unico_libre_de_analfabetismo_en_centroamerica
- Griffin, P. & Care, E. (2015). *Assessment and Teaching of 21st Century Skills: Methods and Approach*. Dordrecht, Netherlands: Springer.
- Hambleton, R. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G. J. Cizek (Ed.). *Setting performance standards: Concepts, methods, and perspectives* (pp. 89–116). USA: Taylor & Francis.
- Hamme, C. & Shulz, M. (2011). Reliability and validity of Bookmark-based methods for standard setting: Comparisons to Angoff-based methods in the national assessment of Educational Progress. *Educational Measurement: Issues and Practice*, 30(2), 3–14
- Hart, C. (1998). *Doing a literature review: Releasing the social science research imagination*. London, UK: Sage Publications.

- Hofstee, B. (1983). The case for compromise in educational selection and grading. In S. B. Anderson & J. S. Helmick (Eds.), *On educational testing* (pp. 109–127). San Francisco: Jossey-Bass.
- Hoyos, C. (2010). *Un modelo para la investigación documental. Guía teórico-práctica sobre construcción de Estados del Arte con importantes reflexiones sobre la investigación*. Medellín: Señal Editora.
- Jaeger, R. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational Measurement* (3rd Ed., pp. 485-514). New York: American Council on Education and Macmillan.
- Janssen, J. & Stoyanov, S. (2012). *Online Consultation of Experts' Views on Digital Competence*. JCR Technical reports. Joint Research Center. European Commission. Recuperado de <http://ipts.jrc.ec.europa.eu/publications/pub.cfm?id=5339>
- Jornet, J. y Gonzales, J. (2009). Evaluación criterial: determinación de estándares de interpretación (EE) para pruebas de rendimiento educativo. *Estudios sobre Educación*, 16, 103-123.
- Kaufman, D., Mann, K., Muijtjens, A. & Van der Vleuten, C. (2000). A comparison of standard-setting procedures for an OSCE in undergraduate medical education. *Academic Medicine*, 75(3), 267-271.
- Leyva, Y. (2011). Una reseña sobre la validez de constructo de pruebas referidas a criterio. *Perfiles Educativos*, 33(131), 131-154.
- Margolis, M. & Clauser, E. (2014). The impact of examinee performance information on judges' cut scores in modified Angoff standard-setting exercises. *Educational Measurement: Issues and Practice*, 33(1), 15–22.
- McClarty, L., Way, D., Porter, C., Beimers, N., & Miles, J. (2013). Evidence-Based Standard Setting: Establishing a Validity Framework for Cut Scores. *Educational Researcher*, 42(2), 78–88. doi.org/10.3102/0013189X12470855
- Ministerio de Educación de Chile. (2013). *Matriz de habilidades TIC para el aprendizaje*. Enlaces: Centro de Educación y Tecnología. Disponible en http://www.enlaces.cl/tp_enlaces/portales/tpe76eb4809f44/uploadImg/File/2015/documentos/HTPA/Matriz-Habilidades-TIC-para-el-Aprendizaje.pdf

- Ministerio de Educación Pública de Costa Rica [MEP]. (2016). *Datos del Curso lectivo 2016*. Dirección de Prensa y Relaciones Públicas. San José: MEP.
- Ministerio de Educación Pública de Costa Rica [MEP]. (2017). *Política Educativa del Siglo XXI*. Recuperado de <http://www.mep.go.cr/politica-educativa>
- Montero, I. y León, O. (2007). A guide for naming research studies in Psychology. *International Journal of Clinical and Health Psychology*, 7(3), 847-862.
- Muñiz, J. (2018). *Introducción a la Psicometría: Teoría Clásica y TRI*. Madrid: Pirámide.
- Muñoz, L., Brenes, M., Bujanda, M., Mora, M., Núñez, O. y Zúñiga, M. (2014). *Las políticas TIC en los sistemas educativos de América Latina: Caso Costa Rica*. Buenos Aires, Argentina: UNICEF.
- Norcini, J. (2003). Setting standards on educational tests. *Medical Education*, 37, 464–469
- Pablos, J., Colás, P., Conde, J. y Reyes, S. (2016). La competencia digital de los estudiantes de educación no universitaria: variables predictivas. *Revista de Pedagogía*, 68, 1-17. DOI: 10.13042/Bordon.2016.48594.
- Popham, J. (1983). *Evaluación basada en criterios*. Madrid, España: Magisterio Español.
- Programa Estado de la Nación [PEN]. (2017). *Sexto informe Estado de la Educación costarricense*. PEN: San José.
- R Core Team (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna: Austria. Recuperado de www.R-project.org/
- Reckase, M. (2006). A conceptual framework for a psychometric theory for standard setting with examples of its use for evaluating the functioning of two standard setting methods. *Educational Measurement: Issues and Practice*, 5(2), 4-18.
- Rodríguez-Bolívar, M., Alcaide-Muñoz, L. & Cobo, J. Analyzing the scientific evolution and impact of e-Participation research in JCR journals using science mapping. *International Journal of Information Management*, 40, 111–119.
- Sanju, G., Sayeed, H. & Femi, O. (2006). Standard setting: Comparison of two methods. *BMC Medical Education*, 6(46), 1-6. doi:10.1186/1472-6920-6-46

- Scharnhorst, A., Börner, K. & Van den Besselaar, P. (2012). *Models of science dynamics: Encounters between complexity theory and information sciences*. Dordrecht: Springer.
- Scolari, C. (2018). *Adolescentes, medios de comunicación y culturas colaborativas. Aprovechando las competencias transmedia de los jóvenes en el aula*. Transliteracy: H2020 Research and Innovation Actions. Barcelona: Universitat Pompeu Fabra
- Sireci, S., Hambleton, R. y Pitoniak, M. (2004). Setting passing scores on licensure examinations using direct consensus. *CLEAR Exam Review*, 15(1), 21-25.
- Sistema de Información de Tendencias Educativas en América Latina [SITEAL]. (2017). *Programa Nacional de Informática Educativa (PRONIE MEP-FOD)*. Recuperado de <http://www.tic.siteal.iipe.unesco.org/politicas/1395/programa-nacional-de-informatica-educativa-pronie-mep-fod>
- Tiffin-Richards, S. & Anand, H. (2013). Setting standards for English foreign language assessment: Methodology, validation, and a degree of arbitrariness. *Educational Measurement: Issues and Practice*, 32(2), 15–25.
- Valdivia, J. (2012). Análisis de pertinencia entre el perfil de egreso y los componentes de la estructura curricular que regulan la actual formación inicial de la carrera de Licenciatura en Educación y Pedagogía en Educación Física de la Universidad de Atacama. *Revista digital Buenos Aires*, 16(164), 1-4.
- Van Eck, N. & Waltman, L. (2018). *Manual for VOSviewer version 1.6.7*. Universiteit Leiden: Netherlands.
- Van Niljen, D. & Janssen, R. (2008). Modeling judgments in the Angoff and contrasting-groups method of standard setting. *Journal of Educational Measurement*, 45(1), 45-63.
- Villalobos, M., Núñez, O., Sequeira, G. y Brenes, M. (2018). Desarrollo de pruebas para evaluar resultados del desempeño de estudiantes costarricenses en el aprendizaje con tecnologías digitales. *Manuscrito pendiente de publicación*. San José: UNED.
- Zieky, M. & Livingston, S. (1977). *Manual for setting standards on the basic skills assessment tests*. Princeton, NJ: Educational Testing Service.

Zieky, M. (2001). So much has changed: How the setting of cut-scores has evolved since the 1980s. In G. J. Cizek (Ed.), *Setting performance standards* (pp. 19-52). Mahwah, NJ: Lawrence Erlbaum.

Zúñiga, M. (2001). *Del construccionismo al constructivismo*. San José, Costa Rica: Fundación Omar Dengo.

VII. Anexos

A. Anexo 1. Fases de evaluación de competencias digitales en la FOD

Fase	Proceso de investigación	Principales productos
I Fase. 2012-2013	Estudio de antecedentes sobre evaluación de este tipo de aprendizajes. Valoración de viabilidad a nivel institucional.	Marco teórico con desarrollo de un marco de indicadores para su evaluación.
II. Fase. 2014	Primera recolección de datos con estudiantes egresados del II ciclo.	Se contó con 9.829 participantes provenientes de 119 colegios ¹¹ .
III Fase. 2015	Primera recolección de datos con estudiantes egresados del III ciclo.	Se contó con 6.767 estudiantes, provenientes de 135 colegios ¹² .
IV Fase. 2016	Segunda evaluación de resultados en estudiantes de II ciclo.	Contando con participación de 20.548 estudiantes ¹³ .
V Fase. 2017-2018	Depuración de banco de ítems y exploración de nuevas metodologías para refinar el tipo de resultado obtenido en el análisis.	Desarrollo de nuevos marcos conceptuales para analizar a e interpretar resultados.

Fases desarrolladas por la Unidad de Investigación de la Fundación Omar Dengo, bajo la supervisión de la directora del Área de Investigación Magaly Zúñiga (Fuente: Elaboración propia a partir de la experiencia como parte del equipo de la Unidad de Evaluación).

¹¹ Detalle de la aplicación documentada en FOD, 2014a.

¹² Detalle de la aplicación documentada en FOD, 2015a.

¹³ Detalle de la aplicación pendiente de publicación en 2017a.

B. Anexo 2. Línea del tiempo de la conformación del PRONIE MEP-FOD

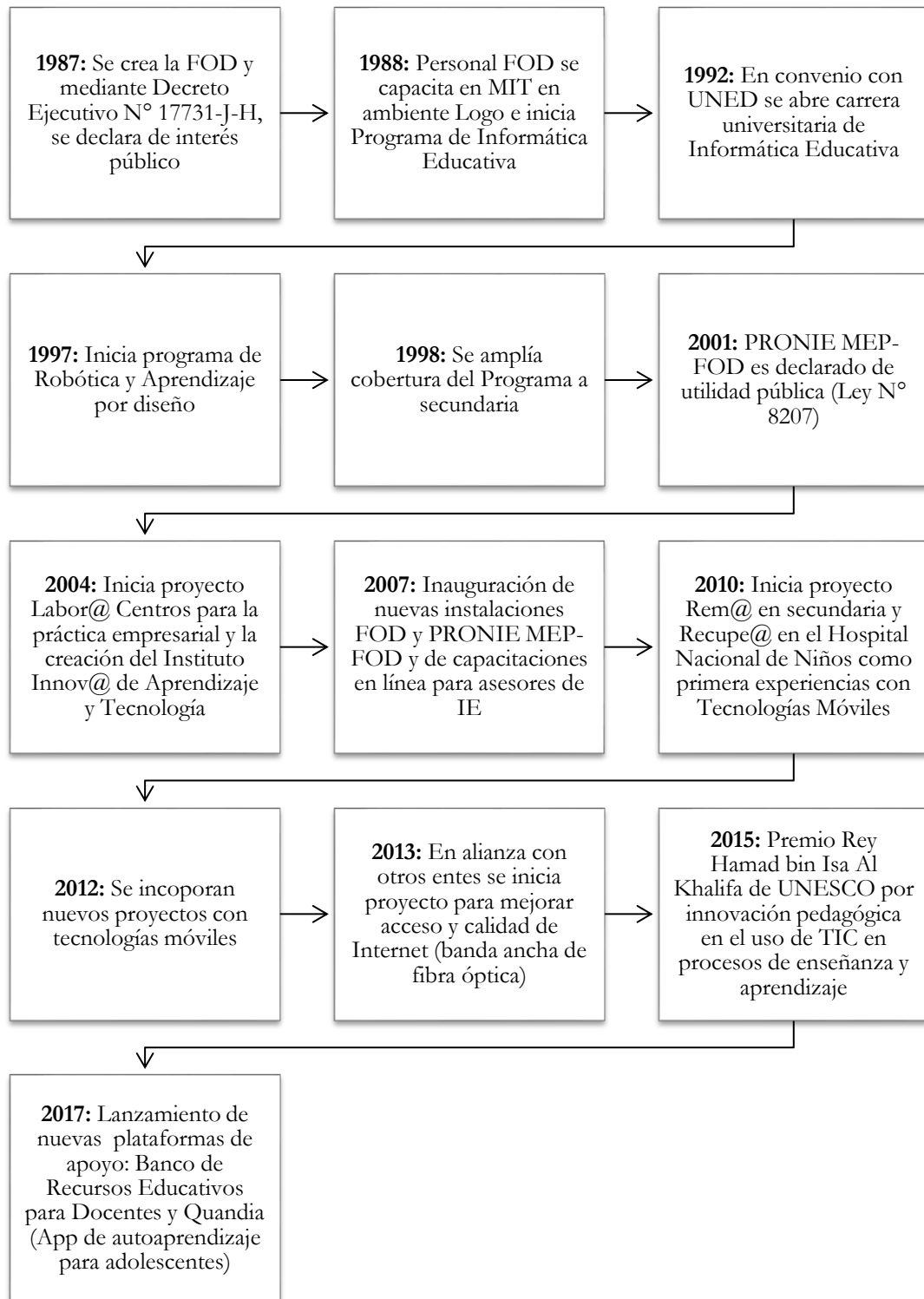


Figura 1. Síntesis de línea del tiempo y principales logros del PRONIE MEP-FOD (Fuente: Elaboración propia a partir de Muñoz et al., 2014 y FOD, 2017b).

C. Anexo 3. Cronograma de trabajo para el desarrollo del TFM

Fecha	Descripción de la actividad	Entregable al cierre de mes
Octubre 2017	Consolidación de la idea de investigación y configuración del equipo tutor.	Documento con ideas iniciales.
Noviembre 2017	Trabajo en delimitación del problema y elaboración del diseño de investigación.	Estructura de documento oficial del diseño de TFM.
Diciembre 2017	Continuación de estudio de documentos para desarrollo de Marco Teórico y comprensión de la temática por investigar.	
Enero 2018	Elaboración del documento oficial de diseño del TFM con las características y requerimientos definidos en la guía.	Capítulo de Introducción: Planteamiento del problema, Revisión de la literatura y Propósito.
Febrero 2018	Diseño de estrategia metodológica, elaboración de materiales requeridos para el establecimiento de los puntos de corte. Así como proyección del tipo de análisis que se hará de los datos.	Capítulo Método: Participantes, Procedimiento, Instrumentos, Diseño y Análisis de datos.
Marzo 2018	Recolección de información del ISI Web of Science y exploración de la herramienta Citespace para llevar a cabo el análisis.	
Abril 2018	Análisis de la información enmarcada en el objetivo específico #1 de investigación.	Avance en capítulo de resultados
Mayo 2018	Análisis de información enmarcada en los objetivos específicos #2 y #3. Redacción de resultados y discusión	Capítulo Resultados Capítulo Discusión
Junio 2018	Redacción final y revisión de aspectos como portada, resumen, índices, reconocimientos, referencias, apéndices.	Primera versión de documento completo
11 de junio 2018	Entrega de TFM a comité asesor previo a defensa	TFM para revisión Presentación para revisión
Del 25 al 29 de junio	Defensa de TFM	TFM versión final Presentación final
10 de julio 2018	Cierre de actas primera convocatoria	

Nota: Se elabora la proyección a partir del calendario definido por la UVa para la elaboración de TFM. Adicionalmente, se toma la *Guía para la elaboración del TFM* del máster para la proyección de los capítulos.

D. Anexo 4. Script ejecutado en R para el análisis bibliométrico

Paso 1. Instalación del paquete bibliometrix

Paso 2. Activación de la biblioteca

```
library(bibliometrix)
```

Paso 3. Asignación del conjunto de datos

```
D <- readFiles("")
```

Ruta establecida para el estudio:

```
D <- readFiles ("C:\\Users\\Melissa\\Documents\\savedrecs1.txt.txt")
```

Paso 4. Conversión del formato de la base de datos

```
M <- convert2df(D, dbsource = "isi", format = "plaintext")
```

```
M
```

Paso 5. Análisis del conjunto de datos

```
results <- biblioAnalysis(M, sep = ";")
```

```
results
```

Paso 6. Tablas resumen de los resultados analizados

```
S=summary(object = results, k = 10, pause =TRUE)
```

Paso 7. Generación de gráficos de tendencias

```
plot(x = results, k = 10, pause = TRUE)
```

Paso 8. Otros análisis ejecutados

Manuscritos más frecuentemente citados:

```
CR <- citations(M, field = "article", sep = ". ") CR$Cited[1:10]
```

Primeros autores más frecuentemente citados:

```
CR <- citations(M, field = "autor", sep = ". ") CR$Cited[1:10]
```

Fuente: Script R_bibliometrix (Aria & Cuccurullo, 2017). Ver más información en <http://www.bibliometrix.org>

E. Anexo 5. Análisis descriptivo del conjunto de datos

Tabla 1. Producción científica anual

Año	Artículos	Año	Artículos
1976	1	2004	22
1982	1	2005	21
1986	1	2006	18
1987	1	2007	19
1990	2	2008	26
1992	7	2009	26
1993	5	2010	24
1994	8	2011	25
1995	9	2012	31
1996	8	2013	40
1997	13	2014	28
1998	4	2015	56
1999	17	2016	53
2000	12	2017	61
2001	15	2018	9
2002	14	<i>Porcentaje de</i>	<i>7,35</i>
2003	24	<i>crecimiento anual (%)</i>	

Nota: Traducción libre de la salida de datos del paquete *Bibliometrix* de R.

Tabla 2. Autores más productivos

Autores	Artículos	Autores	Artículos fraccionados
1 VAN,DERVLEUTENC	20	WYSE,AE	4.75
2 MUIJTJENS,A	10	VAN,DERVLEUTENC	4.44
3 SCHERPBIER,AJJA	9	NORCINI,JJ	3.24
4 NORCINI,JJ	8	PLAKE,BS	2.58
5 PLAKE,BS	7	DOWNING,SM	2.53
6 WYSE,AE	7	KANE,MT	2.17
7 CLAUSER,BE	6	SKAGGS,G	2.17
8 DOWNING,SM	6	MUIJTJENS,A	2.10
9 MARGOLIS,MJ	6	ARNOLD,I	2.00
10 DE,CHAMPLAINA	5	BERK,RA	1.64

Nota: Traducción libre de la salida de datos del paquete *Bibliometrix* de R.

Tabla 3. Países más productivos (correspondiente a los autores)

País	Artículos	Frecuencia
1 USA	310	0,5201
2 ENGLAND	64	0,1074
3 AUSTRALIA	32	0,0537
4 CANADA	27	0,0453
5 NETHERLANDS	25	0,0419
6 GERMANY	21	0,0352
7 NEW ZEALAND	11	0,0185
9 SPAIN	10	0,0168
8 SCOTLAND	9	0,0151
10 TURKEY	6	0,0101

Nota: Traducción libre de la salida de datos del paquete *Bibliometrix* de R.

Tabla 4. Total de citaciones por país

País	Total Citaciones	Promedio de citaciones por artículo
1 USA	5231	16,87
2 ENGLAND	897	14,02
3 NETHERLANDS	735	24,40
4 CANADA	428	15,85
5 AUSTRALIA	331	10,34
6 NEW ZEALAND	166	15,09
7 GERMANY	162	7,71
8 ISRAEL	112	22,40
9 DENMARK	109	36,33
10 IRELAND	99	33,00

Nota: Traducción libre de la salida de datos del paquete *Bibliometrix* de R.

Tabla 5. Manuscritos principales por citación

Artículos	Total de Citaciones (TC)	TC por año
1 ANDERSON JR; CORBETT AT; KOEDINGER KR; PELLETIER R, (1995), J. LEARN. SCI.	642	27.91
2 HELLER P; HOLLABAUGH M, (1992), AM. J. PHYS.	238	9.15
3 RETHANS JJ; NORCINI JJ; BARON-MALDONADO M; BLACKMORE D; JOLLY B; LADUCA T; LEW S; PAGE GG; SOUTHGATE L, (2002), MED. EDUC.	189	11.81
4 CAPELLA J; SMITH S; PHILP A; PUTNAM T; GILBERT C; FRY W; HARVEY E; WRIGHT A; HENDERSON K; BAKER D; RANSON S; REMINE S, (2010), J. SURG. EDUC.	147	18.38
5 NORCINI JJ, (2003), MED. EDUC.	138	9.20
6 MOSS PA, (1992), REV. EDUC. RES.	135	5.19
7 MAYER DP, (1999), EDUC. EVAL. POLICY ANAL.	127	6.68
8 DOWNING SM; TEKIAN A; YUDKOWSKY R, (2006), TEACH. LEARN. MED.	116	9.67
9 KANE MT, (1994), REV. EDUC. RES.	114	4.75
10 JAMSHIDIAN M; BENTLER PM, (1999), J. EDUC. BEHAV. STAT.	111	5.84

Nota: Traducción libre de la salida de datos del paquete *Bibliometrix* de R.

Tabla 6. Manuscritos más frecuentemente citados

Referencias citadas	Frecuencia
1 ANGOFF W H, 1971, ED MEASUREMENT, P508	83
NORCINI JJ, 2003, MED EDUC, V37, P464, DOI 101046/J1365-2923200301495X	30
2 CIZEK G J, 2007, STANDARD SETTING GUI	29
3 LIVINGSTON S A, 1982, PASSING SCORES MANUA	29
DOWNING SM, 2006, TEACH LEARN MED, V18, P50, DOI 101207/S15328015TLM1801_11	27
5 NEDELSKY L, 1954, EDUC PSYCHOL MEAS, V14, P3, DOI 101177/001316445401400101	25
6 MITZEL HC, 2001, SETTING PERFORMANCE STANDARDS, P249	24
NORCINI JJ, 1997, APPL MEAS EDUC, V10, P39, DOI 101207/S15324818AME1001_3	23
8 BEN-DAVID MF, 2000, MED TEACH, V22, P120, DOI 101080/01421590078526	22
9 BERK RA, 1986, REV EDUC RES, V56, P137, DOI 102307/1170289	22

Nota: Traducción libre de la salida de datos del paquete *Bibliometrix* de R

Tabla 7. Palabras clave más relevantes

Palabras clave del autor (DE)	Artículos	Palabras clave-Plus (ID)	Artículos
ASSESSMENT	41	EDUCATION	61
STANDARD SETTING	39	PERFORMANCE	53
STANDARDS	27	STUDENTS	37
EDUCATION	20	TESTS	32
CURRICULUM	15	STANDARDS	27
EDUCATIONAL MEASUREMENT	14	SCORES	22
CLINICAL COMPETENCE	12	COMPETENCE	20
VALIDITY	10	SKILLS	20
REPRODUCIBILITY OF RESULTS	9	VALIDITY	18
SIMULATION	9	CHILDREN	17

Nota: Traducción libre de la salida de datos del paquete *Bibliometrix* de R

F. Anexo 6. Fichas síntesis de las estrategias analizadas

Para cada una de las estrategias identificadas se muestra la valoración de ventajas y desventajas. El detalle de las características específicas y del procedimiento se puede consultar en material digital complementario (matriz de análisis comparativo).

Estrategias centradas en el test

Método Nedelsky

Ventajas y alcances	Desventajas y limitaciones
<ul style="list-style-type: none">- Permite a cada juez dar sus juicios de manera independiente.- Se puede hacer una ronda de discusión para que los jueces revaloren las puntuaciones asignadas (no es obligatorio).- El proceso de obtención del Nedelsky value se realiza de manera sencilla.	<ul style="list-style-type: none">- El valor que puede adquirir cada ítem es muy limitado, considerando que usualmente se trabaja con 4 opciones de respuesta.- Punto de corte binario.- No se recomienda para identificar varios niveles en una prueba, aunque se especifica cómo podría hacerse.- Se suelen obtener puntos de corte más bajos en comparación con los demás métodos debido a la resistencia habitual de los jueces a considerar que todas las personas respondería correctamente el ítem.

2. Método Ebel

Ventajas y alcances	Desventajas y limitaciones
<ul style="list-style-type: none">- Se considera de fácil aplicación.- Los jueces elaboran sus juicios a partir del propósito de la prueba y la población general que se estaría evaluando aunque toma como parámetro hipotético el desempeño de los participantes con menor nivel de habilidad.- Admite una modificación cuando se cuenta con datos empíricos. Consiste en categorizar los ítems según el <i>p value</i> de la siguiente forma: <i>Difícil</i> de 0,00 a 0,49, <i>Medio</i> de 0,50 a 0,79 y <i>Fácil</i> de 0,80 para arriba.- Admite varios tipos de ítems que implican algunas variaciones en la asignación de puntajes de los grupos de ítems, pero no del punto de corte como tal.	<ul style="list-style-type: none">- Al hacer los juicios sobre los ítems resulta difícil para los jueces asignar los dos criterios sin tender a relacionarlos (dificultad y relevancia del ítem), por lo que se recomienda que la asignación de dificultad como se plantea en el método original debería ser sustituida por completo por los datos empíricos.- No se define qué hacer con los ítems que son clasificados por los jueces como "Cuestionables" sino que se utilizan como parte del conjunto de valores en la fórmula para calcular el punto de corte. Esto representa una amenaza para la validez del punto de corte definido y, en general, para la prueba al no excluir o al menos revisar los ítems que el grupo de jueces considera que no responden adecuadamente a la prueba.

3. Método Angoff

Ventajas y alcances	Desventajas y limitaciones
<ul style="list-style-type: none">- Se trata del método más utilizado en la actualidad, por lo que tiene más variaciones en función de los requerimientos que se han ido identificando.- Es uno de los métodos más investigados a profundidad por lo que suele ser incluido como referente en la literatura. Esto conlleva la existencia de amplia bibliografía que compara este método con otros.- Se considera un método fácil de explicar y cuya información y análisis requeridos son más simples que los de otros métodos.	<ul style="list-style-type: none">- Se han hecho muchas variaciones al método original pero se denominan todas como "variaciones al método Angoff" con lo cual se pierde un poco la especificidad del método originalmente planteado.- Una de sus principales limitaciones es que fue originalmente diseñado para ítems de opción múltiple y no es demasiado permeable a la incorporación de ítems con respuesta abierta (<i>constructed response items</i>).- Existe riesgo de que no todos los jueces valoren de igual manera lo que es un sujeto "mínimamente competente", es necesario trabajar de previo en tal conceptualización.- No se ha explorado el funcionamiento de estas variaciones en ítems politómicos.- Se ha criticado como un método defectuoso por partir de una valoración hipotética de sujetos con habilidad límite.- Emite un punto de corte dicotómico (aprueba o no aprueba) por lo que no permite evaluar por niveles de competencia.

4. Variación 1 del método Angoff

Ventajas y alcances	Desventajas y limitaciones
<ul style="list-style-type: none">- Reduce la variabilidad dentro del grupo de jueces utilizando varias rondas de discusión para favorecer consenso.- Amplía el tipo de ítems que se pueden valorar por parte de los jueces.- Cuando se mezclan tipos de ítems se generan puntos de corte para cada tipo, es decir, por componentes que a su vez pueden ser equiparados con métodos más complejos para generar un único punto de corte.- Este método puede adaptarse para definir más de un punto de corte en una prueba, es decir, para generar categorías de desempeño.- Al incorporar información empírica obtenida de aplicaciones previas, esta variación reduce la crítica que se hace al método Angoff de estar basado en un criterio hipotético.	<ul style="list-style-type: none">- Cuando se requiere definir más de un punto de corte se debe someter a los jueces a mayor trabajo y demanda cognitiva al tener que trabajar en la correcta conceptualización de los sujetos límite entre el nivel Básico/Competente y el Competente/Avanzado.- No se ha explorado el funcionamiento de estas variaciones en ítems politómicos como las escalas Likert.- Requiere que los jueces evalúen y estimen los valores p para todos los ítems de la prueba, lo cual suele resultar en una tarea sumamente exigente en cuanto tiempo y recursos cognitivos que implica.

5. Variación 2 del método Angoff (*Angoff Extendido*)

Ventajas y alcances	Desventajas y limitaciones
<ul style="list-style-type: none">- Amplía la posibilidad de aplicación del método al admitir otros tipos de ítems.	<ul style="list-style-type: none">- No se ha explorado el funcionamiento de estas variaciones en ítems politómicos como las escalas Likert.

6. Variación 3 del método Angoff

Ventajas y alcances	Desventajas y limitaciones
<ul style="list-style-type: none">- Simplifica la tarea para los jueces al tener que valorar de manera dicotómica los ítems.	<ul style="list-style-type: none">- Existe un riesgo de que no todos los jueces valoren de igual manera lo que es un sujeto "mínimamente competente" con lo cual resulta necesario trabajar previamente en la conceptualización de esto. Usualmente se requiere generar algún tipo de capacitación.- No se ha explorado el funcionamiento de estas variaciones en ítems politómicos.- Emite un tipo de corte dicotómico.

7. Método de consenso directo

Ventajas y alcances	Desventajas y limitaciones
<ul style="list-style-type: none">- Surge como una alternativa al defecto del método Angoff que implica mucho tiempo por parte de los jueces, así como una mayor dificultad para que éstos hagan el ranking de los ítems.- Se considera más eficiente puesto que reduce significativamente el tiempo que se demanda a los jueces para que hagan la revisión de los ítems, en comparación con otros métodos.- Se promueve el consenso como estrategia de asignación del punto de corte de la prueba, de manera que las razones para definir uno y otro punto son discutidas, argumentadas y valoradas por el grupo completo de jueces.- Se han encontrado menores variaciones entre las calificaciones o valoraciones de los jueces en comparación con otros métodos.- Admite variación de brindar a los jueces información empírica de la prueba aunque no se ha explorado a profundidad el efecto de esto (con los <i>p values</i>).- Se considera que ofrece a los jueces mayor oportunidad de influir en la definición del punto de corte.	<ul style="list-style-type: none">- Si se trabaja con los <i>p values</i> se requiere una importante capacitación de los jueces sobre el uso de esta información ya que se corre el riesgo de que desestimen la consigna inicial de pensar en un sujeto con mínima habilidad y se tienda a valorar los ítems en función del grupo completo.- Sólo funciona con pruebas en las que sea posible subdividir en subáreas el conjunto de ítems (no para pruebas unidimensionales).- Si se trabaja con un amplio grupo de jueces es más difícil generar consenso para definir el punto de corte, con lo cual se debe pensar una alternativa para ello como puede ser trabajar con la media de todos los jueces.- Durante la sesión de consenso es posible que las personas con mayor liderazgo tiendan a tratar de imponer su punto de vista, por lo que es necesario garantizar que haya consenso real integrando todas las opiniones.- Los jueces no reciben información de impacto sobre las consecuencias del punto de corte establecido, con lo cual se recomienda valorar esta posibilidad.

Método de coincidencia en descriptor del ítem

Ventajas y alcances	Desventajas y limitaciones
<ul style="list-style-type: none">- El método es más flexible que otros como el Bookmark en términos de la forma en la que los jueces pueden emitir los juicios sobre los ítems por ejemplo al admitir las regiones límite como una fuente de información dentro del proceso de identificación del punto de corte.- Por las características del método, se recomienda utilizarse en ámbitos educativos que involucran emisión de certificaciones o licencias.- Permitiría generar mapas reales de ítems, que le faciliten al juez visualizar en una sola página el conjunto de ítems que pertenece a cada subárea de la prueba. Sin embargo, a este punto, esta idea se mantiene como una posibilidad a explorar en tanto que no se tiene evidencia de haberse implementado.	<ul style="list-style-type: none">- Considerando que los jueces deben conocer tanto a los evaluados como los contenidos, se reduce la cantidad posible de participantes que se pueden incluir en el proceso.- En algunas experiencias documentadas se ha identificado que los jueces tienen dificultades para ubicar las regiones límite. Se recomienda que los facilitadores de la sesión estén en capacidad de ayudar a los jueces.- Requiere información sobre el desempeño real de los evaluados, es decir, que se necesita contar al menos con una aplicación empírica.- Se requiere ampliar la investigación sobre la determinación y funcionamiento de las reglas con las que se determinan los puntos de corte en las regiones límite.- La validez del método recae en la efectividad y claridad con la que hayan sido definidos los PLD, con lo cual se considera necesario hacer una importante inversión de trabajo en la definición, validación y pilotaje de estos.

Estrategias centradas en el examinado

9. Método de grupo de contraste

Ventajas y alcances	Desventajas y limitaciones
<ul style="list-style-type: none">- El tipo de juicio que se requiere por parte del juez es más sencillo en comparación con el de otros métodos de definición de puntos de corte.- Sólo implica una ronda de valoración por parte de los jueces, con lo cual su implementación puede ser más eficiente en términos del tiempo requerido.- Se propone una variación que implica dividir a los sujetos evaluados por bloques según la calificación obtenida en la prueba y dentro de esa clasificación contabilizar los que fueron ubicados como <i>masters</i> y <i>nonmasters</i>. Siguiendo la lógica de determinación del umbral absoluto, se tomaría como punto de	<ul style="list-style-type: none">- Requiere conocimientos de estadística y software de apoyo como el SPSS.- Cuando se trabaja con muestras pequeñas se recomienda utilizar los modelos de medidas de tendencia central y no el de regresión logística, ya que puede no ajustarse.- Los juicios de los jueces pueden estar más influenciados al sesgo y la inconsistencia por conocer a los sujetos examinados (por ejemplo en cuanto a personalidad, puntualidad, disposición para ayudar, hábitos de trabajo, liderazgo, sexo y etnia, etc).- Requiere la generación de una política de tratamiento de los errores de clasificación,

corte la puntuación que deja por debajo el 50% de los casos categorizados como *masters*.

por ejemplo, con los casos que se comportan como falsos positivos y los falsos negativos.

- No siempre se obtiene el resultado ideal de que todos los sujetos clasificados como *masters* estén por encima del punto de corte y todos los *nonmasters* por debajo, con lo cual se requiere valorar ajustar los puntos de corte para minimizar los errores de clasificación.

10. Método del grupo límite

Ventajas y alcances

- Ofrece a los jueces un punto intermedio para categorizar a los estudiantes, con lo cual se facilita el proceso para quienes tienen dificultad con categorizaciones dicotómicas de tipo *master* o *nonmaster*.
- Ofrece una categorización con juicios ligeramente más finos al tener que categorizar en tres categorías y no en dos como en otros métodos centrados en los examinados.
- Se considera como un método sencillo de implementar siempre y cuando los jueces conozcan bien a los evaluados, de manera que la clasificación que hagan de éstos sean correctas.

Desventajas y limitaciones

- Usualmente la cantidad de sujetos que se clasifican como *borderline* suele ser pequeña, con lo cual el punto de corte puede ser algo más inestable que con otros métodos.
- Se quieren importantes procesos de capacitación para los jueces para evitar errores de tendencia central al clasificar a los sujetos, es decir, que se debe evitar que "por indecisión" se clasifique a todos los sujetos en la categoría intermedia.
- Los juicios de los jueces pueden estar más influenciados al sesgo y la inconsistencia por conocer a los sujetos examinados. Esto podría generar debilidad en términos de confiabilidad y validez.
- Requiere una política de tratamiento de los errores de clasificación (falsos positivos y falsos negativos).

11. Método de captura de juicios políticos

Ventajas y alcances

- Es un método que admite flexibilidad según el perfil que se quiera evaluar.
- Implica un trabajo sencillo por parte de los jueces.
- No solo se enfoca en definir puntos de corte, sino que involucra la articulación de reglas o requisitos para entrar en una profesión.
- Requiere conocimientos básicos de estadística y software de apoyo para asignar los pesos de cada dimensión.

Desventajas y limitaciones

- Es de los métodos holísticos menos utilizados por ser considerado como el más débil a nivel matemático. Esto genera preocupación porque de éste método se derivan políticas por ejemplo de contratación, e incluso, de prioridades en un perfil profesional.
- No se ha estudiado fuera del contexto de certificación de docentes.
- Se recomienda que los jueces sean instructores de los evaluados, con lo cual se reduce la posibilidad de contar con una mayor cantidad de jueces.

12. Método del perfil dominante

Ventajas y alcances	Desventajas y limitaciones
<ul style="list-style-type: none">- Se pueden definir varias políticas de aprobación de la prueba según modelos compensatorios o conjuntivos.- No solo se enfoca en definir puntos de corte, sino que involucra la articulación de reglas o requisitos para entrar en una profesión.	<ul style="list-style-type: none">- Es de los métodos holísticos menos utilizados aunque se considera que va un paso más allá del método anterior.- La variedad de políticas puede representar una amenaza a la confiabilidad del procedimiento. Por lo que se recomienda definir una sola mediante consenso y aplicarla. Sin embargo, no siempre se logra, con lo cual se puede generar confusión a nivel de políticas.- Se recomienda que los jueces sean instructores de los evaluados, con lo cual se reduce la posibilidad de contar con una mayor cantidad de jueces.

13. Método de juicio analítico

Ventajas y alcances	Desventajas y limitaciones
<ul style="list-style-type: none">- Surge como una alternativa a los métodos Angoff y Nedelsky ante la necesidad de integrar nuevos formatos de ítems.- Funciona para definir más de un punto de corte en una prueba.- Se menciona que existen otros métodos propuestos por los autores para calcular el punto de corte; sin embargo, no amplía.- Se propone una variación para simplificar la subcategorización de menos opciones.- Se recomienda para evaluaciones complejas (con distintos formatos de ítems) dentro de los campos de emisión de certificaciones, credenciales y de evaluación educativa.	<ul style="list-style-type: none">- Se ha empleado poco fuera del contexto de la prueba para la que fue creado el método, con lo cual existe poca evidencia empírica sobre su funcionamiento.- Se recomienda que los jueces sean instructores de los evaluados, con lo cual se reduce la posibilidad de contar con una mayor cantidad de jueces.- Se considera que el modelo aun podría ser más simplificado, ya que es complejo. Por ejemplo trabajando únicamente con las tres categorías límite (Límite-básico, Límite-competente y Límite-avanzado).

14. Método de cuerpo de trabajo

Ventajas y alcances	Desventajas y limitaciones
<ul style="list-style-type: none">- Es el más utilizado de los métodos holísticos.- Se considera un método efectivo y eficiente puesto que el juez puede moverse dentro del material con un criterio de ordenamiento previo, lo cual reduce el riesgo de tener inconsistencias en los patrones de juicio.- El método es flexible, puesto que permite reemplazar los conjuntos de muestras que	<ul style="list-style-type: none">- Requiere importante capacitación y práctica previas de los jueces para que comprendan la tarea, se familiaricen con la prueba, con los descriptores de cada nivel de habilidad y que puedan ejecutar efectivamente la tarea (se ejemplifica con una jornada de cuatro días para un conjunto de 26 jueces).

<p>están claramente en un determinado nivel de rendimiento por muestras adicionales con puntajes cercanos a los de mayor desacuerdo.</p> <ul style="list-style-type: none"> - A partir del uso de modelos de regresión logística se reducen errores de medida. - Resulta un método sencillo para jueces como docentes que están habituados a generar "calificaciones" sobre desempeño. - En caso de que el modelo de regresión no se ajustara, se propone como alternativa trabajar con las medidas de tendencia central o, incluso, por votación. 	<ul style="list-style-type: none"> - Se requiere conocimiento de estadística y software (como SPSS, SAS o STATA) por parte de los facilitadores para interpretar los outputs del modelo de regresión logística y valorar (en función de aspectos como el ajuste del modelo, análisis de máxima verosimilitud, entre otros), las decisiones a tomar con respecto al establecimiento de los puntos de corte de la prueba.
---	--

Estrategias de compromiso

15. Método Bookmark	
Ventajas y alcances	Desventajas y limitaciones
<ul style="list-style-type: none"> - Se puede utilizar tanto con ítems de selección única como de producción, con lo cual se amplía la posibilidad de aplicación para pruebas de formatos mixtos. - Implica una tarea relativamente más sencilla para los jueces y con la que, además, podrían estar más familiarizados. Debido a que al organizar primero todo los ítems en orden de dificultad, pueden profundizar el análisis en un rango más estrecho de ítems. - Para los facilitadores también implica una tarea algo más sencilla en términos de preparación previa y de menor tiempo de reunión con los jueces. - En términos psicométricos el método tiene una serie de ventajas porque se basa en los modelos de TRI. - Contar con este modelo permite orientar la construcción de ítems bajo ciertos parámetros que se requieren para aplicar el modelo TRI. Esto favorece mayor calidad en el proceso de construcción de la prueba y, una minimización de los errores de medida. - Permite la valoración de más de un nivel de competencia. - Se ha utilizado en contexto latinoamericano. 	<ul style="list-style-type: none"> - Requiere importante capacitación y práctica previas de los jueces para que comprendan la tarea, se familiaricen con la prueba y con los descriptores de cada nivel de habilidad y lleven a cabo el trabajo requerido. - Se requiere conocimiento estadístico específico en Teorías de Respuesta al Ítem para llevar a cabo el análisis previo del conjunto de datos (cálculo de Response Probability Value o RP según la regla seleccionada). Para ello, también implica el conocimiento y uso de software de apoyo como SPSS, Excel, Winsteps, entre otros. - Existe la preocupación en torno a que el punto de corte es relativo al estar determinado por el grado de dificultad que la prueba tuvo para un conjunto determinado de sujetos. Sin embargo, esta limitación no es exclusiva del método sino que es generalizable a las estrategias de este tipo. - Aunque se plantean algunas posibles variaciones para contrarrestar preocupaciones particulares que han surgido (como usar TC en vez de TRI para identificar el nivel de dificultad -p values u ordenar los ítems de mayor a menor grado de dificultad en el folleto), estas no han sido respaldadas con suficiente material empírico.

16. Método de Hofstee

Ventajas y alcances	Desventajas y limitaciones
<ul style="list-style-type: none">- Se considera uno de los métodos más efectivos en términos de tiempo demandado de parte de los jueces. Se ejemplifica que en una sesión de medio día se puede realizar un entrenamiento adecuado y que la emisión de los juicios no debería superar la hora.- Aunque implica la necesidad de que los jueces estén familiarizados con el contenido evaluado no implica que conozcan a los evaluados.	<ul style="list-style-type: none">- En comparación con otros métodos no se hace una revisión más detallada de los ítems que se utilizaron en la prueba sino que se parte de la premisa de que todos son adecuados y por ello entraron dentro de la prueba.- Se utiliza para determinar un punto de corte binario, es decir, identificar a los estudiantes en términos de aprobados y no aprobados.

Método de Beuk

Ventajas y alcances	Desventajas y limitaciones
<ul style="list-style-type: none">- Es un método sencillo de implementar y cuya demanda de trabajo para los jueces es bastante sencilla en comparación con otros métodos.	<ul style="list-style-type: none">- Se recomienda como método de respaldo o complemento al utilizar otros métodos y no como método principal. Por ejemplo, al finalizar un proceso de establecimiento de puntos de corte con otros métodos, se podrían agregar las cuatro preguntas del método Hofstee o las del Beuk y complementar la información para apoyar la toma de decisiones.- Se generó como método para generar puntos de corte sencillos (aprobó o no aprobó), por lo que no se cuenta con evidencia empírica sobre su uso en más de un nivel de punto de corte (por ejemplo los de categorías como Básico, Competente y Avanzado).- No se cuenta con software específico que permita analizar los resultados de este método, con lo cual su análisis se dificulta un poco más para quienes procesan los datos.- Se han utilizado muy poco, por lo que no se cuenta con suficiente literatura disponible.

G. Anexo 7. Síntesis de sesión de validación de análisis con equipo FOD

Reunión vía Skype: Validación de Standard Setting

Tema: Análisis de pertinencia de las estrategias de definición de puntos de corte identificadas en la literatura.

Fecha: 23 de mayo 2018

Hora de inicio: 14h (hora de CR)

Hora de finalización: 16h

Propósito de la reunión Realizar una validación del análisis de preliminar de pertinencia de las estrategias identificadas en la literatura, según la caracterización del contexto evaluativo de la FOD.

Resultados esperados Categorización final de la pertinencia de las estrategias y selección de la más pertinente para generar propuesta de implementación.

Agenda:

1. Presentar la caracterización de cada uno de los métodos identificados según pautas de interés definidas con equipo de la Unidad de Evaluación de la FOD (tipo de ítem, momento de aplicación, características de los jueces, entre otros).
2. Sintetizar aspectos relacionados con el procedimiento, fortalezas y debilidades generales de cada método.
3. Validar el análisis de pertinencia en función del contexto de evaluación de habilidades en estudiantes del PRONIE MEP-FOD.
4. Orientar la toma de decisiones sobre la estrategia que se perfile como la más pertinente para el contexto de evaluación de la FOD.

Participantes Olmer Núñez y Karol Picado (desde Costa Rica)
Melissa Villalobos (desde España).

Acuerdos

- a. Se elige estrategia Bookmark como la más pertinente para desarrollar la propuesta de implementación.
- b. Olmer y Karol finalizarán la valoración de pertinencia de manera individual y enviarán resultados el día 28 de mayo.
- c. Melissa compilará los comentarios de ambos y con base en el criterio de los tres investigadores actualizará la matriz comparativa con la categorización final de pertinencia para cada estrategia.

Próxima reunión *Tema:* Presentación de resultados del análisis a jefaturas de la Unidad de Evaluación de la FOD.

Fecha: por definir (posterior al 11 de junio)

(Fuente: Elaboración propia a partir de la agenda y ejecución de la sesión).

H. Anexo 8. Ejemplo de plantilla para formulario de respuesta de los jueces

Establecimiento de puntos de corte en evaluación de los Estándares de desempeño

Prueba de evaluación: Aporte de la propuesta de LIE en la promoción de habilidades de Productividad y Resolución de problemas e investigación

Grado evaluado: Sexto grado **Dimensión () evaluada:** () Productividad () Resolución de problemas e investigación

Nombre del juez: _____

Instrucciones:

Por favor, complete los espacios en blanco del siguiente formulario con el número del ítem seleccionado para cada nivel de desempeño. Recuerde que tiene que colocar el número de ítem que se encuentra en la esquina derecha de la hoja del Folleto ordenado de ítems.

Síntesis de los juicios emitidos:

RONDA 1

	Bajo/Medio	Medio/Alto
Nº de Ítem del cuadernillo		

RONDA 2

	Bajo/Medio	Medio/Alto
Nº de Ítem del cuadernillo		

RONDA 3

	Bajo/Medio	Medio/Alto
Nº de Ítem del cuadernillo		

Notas:



(Fuente: Elaboración propia a partir de los ejemplos de Aranguren y Hoszowski, 2017; y Cizek y Bunch, 2007).

I. Anexo 9. Propuesta de agenda de trabajo en sesión con los jueces

Propuesta de agenda para sesión de trabajo con los jueces		
Prueba:	Aporte de la propuesta de LIE en la promoción de habilidades de Productividad y Resolución de problemas e investigación	
Día 1: Contextualización general sobre la evaluación y la estrategia		
Hora	Actividades a desarrollar	Requerimientos
8:00	Registro de los participantes y bienvenida. Instrucciones generales y organización de las sesiones.	Gafetes
8:30	Introducción al trabajo por realizar, distribución de los materiales. Contextualización sobre el proyecto y sobre la tarea por desarrollar. Introducción de conceptos clave como <i>Sujetos con mínimo nivel de habilidad</i> .	Presentación: contexto evaluativo FOD y definición de puntos de corte
10:00	Receso	
10:15	Revisión de la prueba de evaluación: se puede valorar que los jueces la completen y luego se hace discusión sobre ésta.	Prueba aplicada a los estudiantes
12:00	Almuerzo	
13:00	Contextualización sobre los niveles de habilidad esperados. Revisión de las etiquetas y los Descriptores de desempeño preliminares	Presentación: niveles de habilidad esperados
15:00	Comprobación de claridad en las instrucciones y el procedimiento general	Formulario de comprobación de instrucciones
15:30	Finalización de la jornada de trabajo	
Día 2: Capacitación de los jueces en la estrategia		
Hora	Actividades a desarrollar	Requerimientos
8:00	Descripción del procedimiento de implementación de la estrategia Bookmark. Distribución y revisión de materiales de práctica.	Presentación estrategia Bookmark. OIB, formularios y marcadores de práctica
9:45	Receso	
10:00	Práctica sobre el establecimiento de puntos de corte con datos de ejemplo. Trabajo en subgrupos.	OIB, formularios y marcadores de práctica
11:15	Discusión grupal de dudas y realimentación sobre el proceso.	
12:00	Almuerzo	
13:00	Instrucciones para Ronda 1 de establecimiento de puntos de corte y distribución de material.	Presentación con instrucciones. OIB,

formularios y marcadores
oficiales

- 13:30 Ronda 1 (Oficial). Trabajo en subgrupos para ambas dimensiones
- 15:00 Realimentación sobre el proceso y cierre de la sesión
- 15:30 Finalización de la jornada de trabajo

Día 3: Obtención de los puntos de corte oficiales

Hora	Actividades a desarrollar	Requerimientos
8:00	Distribución de materiales y realimentación sobre Ronda 1 de valoraciones (información normativa).	Información normativa Ronda 1
8:30	Ronda 2. Trabajo en subgrupos para ambas dimensiones	
10:00	Receso	
10:15	Distribución de materiales y realimentación sobre Ronda 2 de valoraciones (información normativa y de impacto).	Información normativa y de impacto Ronda 2
10:45	Ronda 3. Trabajo en subgrupos para ambas dimensiones	
12:00	Almuerzo	
13:00	Discusión global sobre los resultados y revisión de los Descriptores de cada nivel.	Presentación con puntos de corte obtenidos y descriptores por nivel para cada dimensión
14:30	Recomendaciones finales del grupo de jueces sobre la estrategia	
15:00	Evaluación final de la jornada	Formulario de evaluación
15:30	Agradecimiento y finalización del proceso	

Notas:



(Fuente: Adaptación del modelo de Cizek y Bunch, 2007, p. 178).

J. Anexo 10. Escala de valoración de la sesión por parte de los jueces

Establecimiento de puntos de corte en evaluación de los Estándares de desempeño		
Prueba de evaluación:	Aporte de la propuesta de LIE en la promoción de habilidades de Productividad y Resolución de problemas e investigación	
Instrucciones: A continuación se presenta una serie de afirmaciones respecto de las tareas que ha realizado en los últimos tres días. Por favor, lea cada frase y señale si está “De acuerdo” o “En desacuerdo”. Al terminar puede agregar aquellos comentarios que considere podrían enriquecer este proceso a futuro.		
	De acuerdo	En desacuerdo
1. Durante la capacitación pude entender claramente el propósito de las tareas a realizar	()	()
2. Los asistentes y coordinadores de las jornadas explicaron claramente las actividades que debíamos realizar	()	()
3. La capacitación y la ronda práctica me ayudó a entender cómo realizar las tareas de cada ronda	()	()
4. Revisar los ítems del cuadernillo me ayudó a comprender mejor la evaluación	()	()
5. Las etiquetas de los niveles me parecieron claras y útiles	()	()
6. Las discusiones en los grupos pequeños y en el grupo completo me ayudaron a comprender el procedimiento y realizar la tarea con mayor eficacia	()	()
7. El tiempo previsto para las rondas de discusiones fue adecuado	()	()
8. Todos los participantes del pequeño grupo en el que participaba tuvieron las mismas oportunidades para expresar sus opiniones y puntos de vista	()	()
9. Fui capaz de seguir las instrucciones y completar el formulario en cada ronda	()	()
10. Las discusiones realizadas después de la primera ronda me resultaron útiles	()	()
11. Las discusiones realizadas después de la segunda ronda me resultaron útiles	()	()
12. Los datos brindados en las devoluciones de cada ronda por los asistentes y coordinadores me resultaron útiles para realizar la tarea	()	()
13. Me siento seguro de los marcadores establecidos y creo que son apropiados para la población implicada	()	()

14. Las facilidades de alojamiento, viáticos y servicios brindados () ()
por los coordinadores generaron un buen clima de trabajo

Notas:



(Fuente: Adaptación del modelo de Aranguren y Hoszowski, 2017, p. 52).

“El viaje no termina jamás. Solo los viajeros terminan. Y también ellos pueden subsistir en memoria, en recuerdo, en narración... El objetivo de un viaje es sólo el inicio de otro viaje” — J. Saramago