Análisis de validez de la prueba Planea de EMS 2017, área matemática

M. A. León Pérez, J. J. Báez Rojas y M. G. Corona-Galindo
Instituto Nacional de Astrofísica, Óptica y Electrónica
Tonantzintla, Puebla, México
a01002417@gmail.com; jjbaezr@inaoep.mx; mcorona@inaoep.mx

Resumen. Las pruebas estandarizadas son elaboradas con el fin de hacer comparaciones objetivas entre alumnos en diversos contextos y son aplicadas a gran escala. La validez, confiabilidad, facilidad en la aplicación y calificación son características necesarias para que una prueba sea objetiva. La validez es la condición que se cumple cuando una prueba mide aquello para lo que fue creada; para lograrlo, cada ítem debe tener una respuesta unívoca e inconfundible, escrita en un lenguaje claro, preciso y acorde al nivel y capacidad del sustentante. De ahí la pertinencia del presente trabajo; pues, el análisis que se hace sugiere que siendo el examen Plan Nacional para la Evaluación de los Aprendizajes (Planea) un instrumento que evalúa a todo un país no debería tener errores en la construcción de la prueba ni en sus reactivos. Con base en un conjunto de directrices, se propone una matriz para calificar la validez de cualquier test de opción múltiple y se aplica a la prueba Planea parea Educación Media Superior del 2017 en el área de matemáticas, encontrándose que el 10% de los ítems de dicha prueba tiene errores.

Palabras clave: pruebas estandarizadas, validez, reactivos de opción múltiple, matriz, prueba Planea.

Abstract. The standardized tests are developed for making objective comparisons between students in different contexts and they are applied on a large scale. The validity, reliability, application-ease, and qualification are necessary characteristics for a test to be objective. Validity is the condition that is fully satisfied when a test measures that for which it was designed; to achieve this, each item must have a unique and unmistakable answer, besides, it must be written in a clear, precise language and in accordance with the level and capacity of the student which takes the exam. Hence the relevance of the present work; since, the analysis that it is carried out, suggests that being the Plan Nacional para la Evaluación de los Aprendizajes (Planea) an instrument which evaluates a whole country, it should not have errors in the construction of the test nor in its reagents. On the basis of a set of guidelines, a matrix is proposed to qualify the validity of any multiple-choice test and it is applied to the Planea school examination in 2017 on the mathematics area only, finding that 10% of the items of such test have errors.

Keywords: standardized tests, multiple choice items, matrix, validity, Planea test.

Análisis de validez de la prueba Planea de EMS 2017, área matemáticas

1. Introducción

La prueba Plan Nacional para la Evaluación de los Aprendizajes de Educación Media Superior (Planea EMS) y el Programa Internacional de Evaluación de los Alumnos (PISA, por sus siglas en inglés) son pruebas que pretenden dar cuenta del aprovechamiento de los estudiantes de Educación Media Superior (EMS) en lo que concierne a dos áreas de competencia, a saber Lenguaje y Comunicación (Comprensión Lectora) y Matemáticas. El formato que gobierna dicha prueba es el de Reactivos de Opción Múltiple (ROM).

En México, El Centro Nacional de Evaluación para la Educación Superior (CENEVAL) es la institución encargada de elaborar los reactivos para las pruebas Planea y para aplicarlas se dió a la tarea de elaborar un Manual de Elaboración de Reactivos (SEP, 2016), donde establece los lineamientos para la elaboración de los ítems y la manera de contestarlos, comenzando por cimentar la exigencia de que para resolver los ejercicios no debe ser necesario el uso de calculadoras ni de fórmulas especializadas.

La elaboración de este tipo de pruebas es un gran desafío para los investigadores de la educación; así, por ejemplo, Haladyna, Downing y Rodríguez (2002) sintetizan más de cuarenta taxonomías que existían hasta los inicios del presente siglo y la dividen en cinco pautas en donde hacen propuestas referentes al contenido, formato, estilo, lineamientos para la redacción del enunciado y la redacción de las opciones de respuesta. La organizan, además, en 31 epígrafes que son básicamente recomendaciones para la elaboración de reactivos coherentes; una adecuación de ésta es la presentada en la Tabla 2, la cual contiene 10 pautas: sobre el contenido, sobre la base, las opciones de respuesta, la respuesta correcta, los distractores, las imágenes, los criterios editoriales, las citas y unidades de medida y la puntuación. Los epígrafes ascienden a 52.

Por otro lado, el trabajo de Moreno, Martínez, y Muñiz, (2004), permite vislumbrar nuevas directrices como resultado de una filtración realizada a la taxonomía de Haladyna (2002). Los autores presentan tres pautas con 12 epígrafes, *videlicet*, elección del contenido que se desea evaluar, expresión del contenido en el ítem y construcción de las opciones la cual puede consultarse en la Tabla 4. Estas propuestas adolecen de un análisis de validez a pruebas estandarizadas a gran escala y tal concepto es un criterio de calidad relacionado con la adecuación de las puntuaciones del test y el objetivo que suscitó su aplicación. Al respecto Arias (1995) establece diferentes formas de calificar la validez y Messick (1989) apunta que también se deben tomar en cuenta las puntaciones obtenidas en la prueba para poder calificarla adecuadamente. Un referente internacional de evaluación educativa, el de los *Standards for Educational and Psychological Testing* establece que la validez, se refiere al grado en el que la evidencia y la teoría apoyan las interpretaciones de los resultados de una prueba para los usos propuestos del examen.

1.1 El contexto de la evaluación de la Educación Media Superior en México

A partir del año 2015 la versión de la prueba Evaluación Nacional de Logro Académico en Centros Escolares (ENLACE) para la Educación Media Superior, se transformó en Planea EMS y para el año 2017 se aplicó los días 4 y 5 de abril de 2017. El propósito de dicho examen es conocer en qué medida los estudiantes logran dominar un conjunto de aprendizajes esenciales al término de la EMS, en dos áreas de competencia: Lenguaje y Comunicación (Comprensión Lectora) y Matemáticas. En el presente trabajo nos restringiremos al área de matemáticas.

La prueba Planea EMS retoma en su contenido matemático las competencias básicas disciplinares sobre las que se basa para detectar si el sustentante las ha desarrollado suficientemente, enfocándose a tres

contenidos matemáticos: cantidad, espacio y forma, y cambios y relaciones. Para complementar, los procesos cognitivos que el sustentante debe dominar son tres: reproducción, conexión y reflexión (SEP, 2016). Sin embargo, lo anterior fue cambiado para la versión del examen Planea EMS 2017 (SEP, 2017b), ya que se plantea que los aprendizajes clave a evaluar tienen los siguientes ejes temáticos los cuales se marcan en el Nuevo Modelo Educativo (SEP, 2017c):

Ejes	No. reactivos
Sentido numérico y pensamiento algebraico	18
Cambios y relaciones	17
Forma, espacio y medida	5
Manejo de la información	10
Total	50

En el presente año y en el área de Matemáticas, la prueba mencionada se aplicó a los alumnos del último ciclo de la EMS y en su propio lenguaje cita que "es una evaluación diagnóstica individual que consta de 50 reactivos de opción múltiple, es una prueba objetiva y estandarizada. No tiene un correlato directo con contenidos curriculares o prácticas de enseñanza que pudieran estar determinadas en los planes y programas de estudio" (SEP, 2017a).

Los resultados de las pruebas Planea EMS y PISA son publicados nacional e internacionalmente colocando al país por debajo de los resultados aprobatorios de dichas evaluaciones, según los resultados publicados en octubre de 2017 sobre Planea EMS, "en Matemáticas, 6 de cada 10 estudiantes se ubica en el nivel I (66%); casi 2 de cada 10 se ubican en el nivel II (23 %); en el nivel III, sólo 8 de cada 100 estudiantes (8%); en el nivel IV, casi 3 estudiantes de cada 100 (2.5%)", tal como se muestra en la Figura 1. De manera paralela solo que, teniendo como panorama los resultados correspondientes al informe de Resultados de PISA 2015 (OCDE, 2016), el desempeño de México (408 puntos) se encuentra por debajo del promedio de todos los países de la OCDE (490 puntos). Por si esto fuera poco, nuestra nación se ubica por debajo de países latinoamericanos como Chile (423 puntos) y Uruguay (418 puntos), lo cual muestra que estamos peor en desempeño en matemáticas que estos dos países. En contraste, estamos ligeramente mejor que Brasil (377 puntos), Colombia (390 puntos), República Dominicana (328 puntos) y Perú (387 puntos).

Sin embargo, aun cuando México tiene calificaciones por arriba de algunos países de la OCDE no es garantía de tener buenos niveles en desempeño educativo. En promedio, en los países de la OCDE, cerca de uno de cada diez estudiantes alcanzan un nivel de excelencia (10.7%) En contraste, en México, un estudiante de treinta alcanza el nivel de excelencia (0.3%) en competencia en matemáticas. De los países de la OCDE casi uno de cada cuatro estudiantes (23%) no al alcanza el nivel básico de competencia, mientras que en México el 57% de los estudiantes no alcanzaron el nivel básico de competencias. Aún más, los estudiantes que no alcanzan el nivel básico de competencias pueden de vez en cuando realizar procedimientos rutinarios canónicos, tales como operaciones aritméticas en situaciones donde todas las instrucciones se les proveen, pero no están facultados para representar matemáticamente una simple situación del mundo real; por ejemplo, comparar la distancia total entre dos rutas alternativas, o convertir precios a una moneda diferente (OCDE, 2016).

En una prueba de opción múltiple, la tarea más importante en su elaboración es la construcción de los ítems, pues según Osterlind (1998) cada uno de ellos es "una unidad de medida en una prueba aplicada que consta de un estímulo y una forma prescriptiva de respuesta y su fin es inferir la capacidad del

examinado en un cierto constructo (habilidad, rasgo, etc.), proporcionando datos cuantificables sobre la persona que lo completa".

Para el caso específico de la prueba Planea EMS aplicada en 2017 la elaboración se basa en el Manual de Elaboración de Reactivos (SEP, 2016) y un análisis de dicho manual muestra ser una adecuación de la metodología propuesta por Haladyna (2002) sin tomar en cuenta el contexto nacional de instalaciones, acceso a la información y preparación del docente. Minucias aparentemente insignificantes, pero que el Instituto Nacional para la Evaluación de la Educación (INEE) ya reconoce, citando "que los resultados son un reflejo de múltiples factores, desde las actividades escolares de los estudiantes (hábitos, actitudes y valores), hasta las condiciones de las instituciones educativas y el contexto socioeconómico en el que viven" (INEE, 2017); sin embargo parte de estos resultados también son resultado del poco cuidado que se tuvo en la elaboración de los reactivos que conformaron dicha prueba.

Con este telón de fondo, el propósito y objetivo del presente trabajo es proponer un instrumento que permita aportar un camino sencillo y rápido para evaluar la validez de ROM y en particular aplicarlo a la prueba Planea EMS 2017. A este instrumento le dimos una forma matricial, pero, antes de adentrarnos en los detalles de la elaboración de esta matriz, nos centraremos en algunos de los estudios más representativos que se han llevado a cabo en esta línea de investigación. Haladyna (2002) propone indicadores sobre la evidencia de validez relacionada con el contenido para elaborar reactivos de opción múltiple, además de reportar que analizó más de un 90% de las taxonomías existentes hasta el año el 2001. En un estudio posterior Moreno, Martínez y Muñiz (2004), presentaron una propuesta para la elaboración de ROM, aportando evidencias de validez relacionadas con el contenido. En este trabajo se dejó de lado, por el momento, el trabajo de Rivera, Flores, Alpuche y Martínez (2016) porque en su propuesta prioriza la evaluación de los resultados a través de los *Standards for Educational and Psychological Testing* y este análisis es el objetivo de un trabajo posterior.

1.2 Justificación de la presente investigación

Debido a los bajos resultados en aprovechamiento que ha mostrado el estudiantado mexicano en las pruebas PISA y Planea EMS y siguiendo con la política gubernamental de trabajar en la preparación de nuestros estudiantes para la superación de dichas pruebas y no haber logrado éxito alguno -pues aunque se ha obtenido una mejor calificación ésta aún no ha sido aprobatoria- sirvió de acicate para buscar las metodologías de elaboración de ROM que rigen dichas pruebas para, posteriormente, analizar la prueba Planea EMS, que es la vigente para calificar el aprovechamiento de los estudiantes en el rubro de Comprensión Lectora y Matemáticas. Restringiendo el análisis sólo a matemáticas, se consideró necesario construir una taxonomía para la elaboración de ítems de opción múltiple que empatara con los programas de estudio vigentes y por añadidura nos permitiera analizar la validez de la prueba Planea EMS. En este contexto nació la taxonomía que se pone a consideración.

1.3 Referencias conceptuales

Una prueba objetiva, ya sea estandarizada o no estandarizada, debe tener tres características: validez, confiabilidad, facilidad en la aplicación y calificación. La validez es un criterio de calidad relacionado con la adecuación de las puntuaciones del test y el objetivo que suscitó su aplicación; además, la validez es un concepto unitario, aunque según Arias (1995) hay diferentes formas de recoger evidencias del concepto como validez de constructo, contenido, criterio. etc. Asimismo, la validez suele entenderse con relación al uso que se dé al cuestionario y la interpretación de sus puntuaciones; sin embargo, Messick (1989) cita

que no se debe validar sólo el cuestionario, sino también las inferencias o interpretaciones que se hagan a partir de las puntuaciones obtenidas con el mismo. Tristán y Pedraza (2017) arguyen que la objetividad junto con la validez y la confiabilidad, configuran los tres atributos fundamentales para el diseño, administración, interpretación y uso de las pruebas estandarizadas.

Respecto a la validez, Rodríguez (1999) apunta que "es la condición que se cumple cuando una prueba mide aquello para lo que fue creada". Para lograrlo, debe ser:

- **Unívoca:** Cada pregunta o reactivo ha de tener una sola respuesta precisa e inconfundible.
- **Inequívoca:** Su lenguaje debe ser tan claro y preciso que evite interpretaciones falsas.
- **Adaptada:** De acuerdo con los métodos y programas de las materias, en correspondencia con el nuevo currículo, así como al nivel y capacidad de alumno.
- **Suficiente:** Ha de tener todos aquellos aspectos considerados como fundamentales.
- **Económica:** En su resolución se ha de emplear el tiempo prudente sin alargarla innecesariamente.

Para los autores, el concepto de validez de una prueba contiene varias aristas: la primera es que el reactivo esté gramatical y semánticamente bien planteado y la relación pregunta-respuesta sea biunívoca. Además, que el reactivo sea acorde con los contenidos que se pretende evaluar y que haya coherencia entre la calificación obtenida y el conocimiento real del estudiante.

1.4 Metodología

Se comenzó con una revisión bibliográfica sobre la elaboración de los siguientes rubros: elaboración de ROM, validez, las taxonomías existentes para la elaboración de ROM, pruebas estandarizadas a gran escala, *viz* PISA y Planea. Posteriores, se juzgó la pertinencia, para el caso de México, de dichas taxonomías y se tamizaron las de Haladyna (Tabla 2), Moreno (Tabla 4), y CENEVAL (Tabla 3) para crear la de León_Báez_Corona (Tabla 1). La pertinencia se determinó considerando los programas de estudios de EMS y básica, tomando como elemento de evaluación la prueba Planea EMS 2017, enmarcando el trinomio en el concepto de validez definido anteriormente. Con estas consideraciones, se elaboró la matriz con las nueve directrices que conforman la taxonomía de los autores del presente trabajo.

1.4 Contexto de la elaboración de pruebas de opción múltiple

En la revisión bibliográfica que se llevó a cabo se encontraron pocas referencias vinculadas al tema que nos ocupa; pues la mayor parte se enfoca a la elaboración de ítems de opción múltiple centrándose en la redacción; la forma de presentar las preguntas y sus componentes, pero nadie enfocó su análisis a la evaluación de su validez, desde la perspectiva de la estructura del examen y la especificación de los reactivos. La técnica de opción múltiple consiste en plantear una pregunta a contestar o un problema a resolver con varias opciones de respuesta, en ambos casos; de las que la selección de respuesta es única. Con el mismo formato de pregunta y respuesta, también puede contener afirmaciones a valorar, característica a cubrir o acción a realizar. A propósito de cómo elaborar un ROM, Haladyna (2002) sintetizan más de cuarenta taxonomías que existían hasta los inicios del presente siglo, dividida en cinco pautas y organizada en 31 epígrafes, tal como se muestran en la Tabla 2.

El Centro Nacional de Evaluación para la Educación Superior enmarca la prueba en las 8 competencias del Marco Curricular Común (MCC) de donde se seleccionaron las 6 siguientes (Acuerdo Secretarial 444, 2008):

1. Interpreta modelos matemáticos mediante la aplicación de procedimientos aritméticos, algebraicos, geométricos y variacionales, para la comprensión y análisis de situaciones reales, hipotéticas o formales.
2. Resuelve problemas matemáticos, aplicando diferentes enfoques.
3. Interpreta los datos obtenidos mediante procedimientos matemáticos y los contrasta con modelos establecidos o situaciones reales.
4. Analiza las relaciones entre dos o más variables de un proceso social o natural para determinar o aproximar su comportamiento.
5. Cuantifica y representa matemáticamente las magnitudes del espacio y las propiedades físicas de los objetos que lo rodean.
6. Lee tablas, gráficas, mapas, diagramas y textos con símbolos matemáticos y científicos.

Por otro lado, el trabajo reportado por Moreno et al. (2004), permite vislumbrar nuevas directrices como resultado de una filtración realizada a la metodología de Haladyna et al. (2002) obteniendo 12 directrices que se pueden ver en la Tabla 4. Estos autores justifican en el apartado "Procedimientos para aumentar la parsimonia" la nueva versión en la que se excluyen directrices de la original de Haladyna (2002). La versión de Moreno sobre la exclusión de algunas directrices de Haladyna se basa en la ambigüedad existente entre ellas, la dependencia que las vincula y ausencia de integraciones que pudieran aportar mayor parsimonia.

2. Propuesta del problema de investigación: Elaboración de una taxonomía para la elaboración de pruebas con ítems válidos de opción múltiple y aplicación de ella para analizar la prueba Planea EMS 2017

Tomando como base las 31 directrices establecidas por Haladyna et al. (2002), tamizadas por el trabajo de Moreno et al. (2004) y considerando también la metodología usada por CENEVAL para la elaboración de ROM, llegamos a la conclusión de que son suficientes 9 directrices compactas para hacer un análisis de validez de los ítems de opción múltiple. Estas directrices dieron pie a una matriz que permite evaluar los reactivos de forma independiente y objetiva la cual puede consultarse en la liga http://astro.inaoep.mx/archivos/nathalie/Corona/Matriz_Leon_Baez_Corona.pdf

Usar una matriz como herramienta de trabajo para tamizar un examen de un ROM fue de gran utilidad ya que como lo establece Rodríguez (1999), en su glosario define una matriz o tabla de especificaciones como tablas que pueden definir el contenido de una prueba; son particularmente útiles para la elaboración de pruebas de conocimiento y para determinar su validez. Usualmente se establecen como tablas de doble entrada (aunque pueden incluir más de dos dimensiones), una de las cuales indica el contenido del proceso de aprendizaje y la otra alguna definición de las habilidades que la persona alcanzadas durante ese proceso.

El conjunto de 9 directrices que integran nuestra propuesta permite en general evaluar ítems de opción múltiple y en particular los de Planea EMS 2017. Dichos criterios son independientes entre sí y se presentan sintetizadas en la Tabla 1. Estas directrices están centradas en solamente cuatro opciones de respuesta, pues son las que CENEVAL establece, en donde, además, la respuesta correcta debe estar colocada en cualquier lugar entre las distintas ubicaciones posibles para que estas no aporten información

indebida; tal como, que el sustentante conteste todo el examen con una sola opción, o bien, contestando en una secuencia ascendente o descendente y acierte en más del 25%. Las directrices que establecemos son:

A. Contenido a evaluar

1. El ítem presenta un solo contenido temático.
2. Los contenidos de la prueba PLANEA 2017 deben ser parte de la currícula de la EMS o bien recordar habilidades y conocimientos adquiridos en la Educación Básica como lo establece el Manual para Usuarios 2016.

B. Expresión del contenido en el ítem

3. El tallo del ítem plantea la idea central, haciendo de cada opción un complemento que debe concordar sintáctica y semánticamente.
4. La sintaxis o estructura gramatical utilizada debe ser correcta, de tal manera que el sustentante se concentre en la pregunta y no en cuestiones gramaticales; evitar un ítem demasiado escueto o excesivamente profuso en su redacción, ambiguo o confuso, no contener oraciones negativas porque pueden resultar complicadas de entender. La semántica debe ser acorde al contenido y a la población de sujetos evaluados, para que pueda ser comprendida sin dificultad.

C. Expresión de las opciones

5. La opción correcta de cada ítem debe ser sólo una, acompañada de opciones plausibles
6. Cada opción debe ser un complemento de la pregunta y concordar sintáctica y semánticamente. No usar términos demasiado inclusivos o excluyentes como: **nunca**, **siempre**, **nadie, todos**, **único**. Evitar expresiones que representen la incapacidad de pensar en otras posibilidades: **ninguna de las anteriores**, **o todas las anteriores**.
7. Las opciones deben presentarse en vertical, para facilitar su lectura, esta distribución permite destacar espacialmente el ítem y facilitar su lectura; así como su diferenciación del resto de los que conforman una prueba, tal como lo establece el Manual para Usuarios 2016 que hizo CENEVAL (SEP, 2016).
8. Ninguna de las opciones debe destacar del resto por ser la única diferente en contenido, en aspectos de apariencia como longitud, estructura gramatical o en algún término que aporte información indebida.
9. El conjunto de opciones de cada ítem debe ser organizado u ordenado de modo coherente con el contenido en estudio y enmarcado en el concepto de validez definido en la metodología.

Teniendo en cuenta la importancia de mostrar al lector los conceptos delineados anteriormente de una manera concisa, a fin de que los comprenda y los pueda aplicar al nivel o escuela de su competencia, así como también a cualquier área de conocimiento, se sintetizó la taxonomía en la Tabla 1.

Desde la perspectiva de Oliden (2003), el análisis de validez de una prueba se debe hacer respecto a la construcción de la misma, la recopilación de evidencias e interpretación de las puntuaciones.

Por lo que concierne a la validez parcial de la prueba, nuestro instrumento contiene, además de las 9 directrices, tres apartados que no ofrecen puntuación alguna, pues no existen razones a priori para asignarles peso de medida y están marcadas en la matriz por un pulgar verde apuntando hacia arriba (***like)*** y un pulgar rojo apuntando hacia abajo (***unlike***), Esto con la finalidad de tener una rápida visualización de aquellos ítems que cumplen, en el primer caso, con las directrices especificadas en la matriz y, en el segundo, aquellas que no las satisfacen.

Por lo que respecta al análisis, se insertaron por separado la pregunta y las opciones de respuesta. Estas últimas se enmarcaron con amarillo para destacarlas del resto del texto. Quedan pendientes dos aspectos de evaluación; a saber, las evidencias: cuántos estudiantes contestaron bien el ítem, y cuáles fueron las puntuaciones finales que obtuvieron en el examen. Hacer el análisis estadístico de estas dos categorías de evaluación es el contenido de dos artículos posteriores que están en preparación, pues la Secretaria de Educación Pública (SEP b) recién dio a conocer los resultados.

Debido a que ningún instrumento de evaluación, por más amplio que sea, puede comprender un examen completo de conocimientos, el análisis de cada uno de los ítems permite comprobar que no todos los contenidos del área de matemáticas del nivel medio superior están considerados. Asimismo, se encontraron ítems que eran parte del mismo contenido de un mismo bloque, lo cual conlleva al desaprovechamiento de los espacios y oportunidades para incluir otros bloques dignos de evaluación.

La matriz permite consultar el contenido de cada ítem y ligarlo a los programas de estudio de secundaria y EMS a nivel federal. En los ítems se marcan con círculos rojos aquellos errores detectados que contravienen los lineamientos previamente establecidos por el propio CENEVAL, e.g. *L* para litros, habiéndose establecido *l*. La última directriz de la matriz versa sobre el orden: de acuerdo a los contenidos de estudio, la mayor parte de ellos se encuentra disperso, ya que algunos de menor orden en los contenidos se encuentran por arriba de los de mayor orden.

3. Resultados y Discusión.

Consideramos que la falta de evaluación de los ítems de la prueba ENLACE en años anteriores, se reflejó en las copias que de ella hizo Planea EMS en los años 2015 y 2016. Esto se debe en gran medida a que, aun cuando se cuenta con un Manual de Elaboración de Reactivos por parte de CENEVAL, no se hizo una revisión a posteriori de la impresión de la validez de cada uno de los ítems que permitiera evaluarlos adecuadamente en algún marco preestablecido, como el que se presenta, por ejemplo. A fin de coadyuvar en tal empresa, es pertinente la matriz con pocas directrices que se propone en el presente trabajo, ya que abre incluso la posibilidad de que el público versado en evaluación pueda emitir un juicio objetivo sobre la validez de la prueba que nos ocupa. Aun cuando CENEVAL define el formato que debe caracterizar cada ítem se han encontrado, errores en su construcción, tal como puede observarse en la Figura 2. Esto obliga a revisar cada reactivo y modificarlo; ya que en consonancia con Adam y Caldwell (2014) algunas violaciones a estos criterios pueden modificar el comportamiento psicométrico del ítem. Un análisis de los resultados que recién se publicaron serviría para confirmar que la detección de cualquier falla en la elaboración de un reactivo de opción múltiple tiene repercusiones incalificables.

Con base en las referencias a la que se tuvo acceso, los exámenes a gran escala estructurados con ítems de opción múltiple constituyen una técnica útil; sin embargo, el diseño de buenos exámenes debe tener un instrumento de evaluación de fácil acceso que permita tener una evidencia rápida de su validez. Además, es importante que los ítems sean diseñados por profesionales que conozcan la currícula de la EMS,

demanda que exige con apremio la creación permanentemente de cursos-taller a fin de aprender a elaborar ítems sobresalientes.

Consideramos, que otro factor que debe tenerse en cuenta, para la buena elaboración de una prueba aplicable a todo el país, es iniciar un estudio del contexto nacional sobre la materia a evaluar, considerando, además, las instituciones educativas y sus instalaciones, así como al magisterio con su nivel de preparación y calidad. Esto coadyuvaría a tener elementos previos suficientes que permitieran la construcción de ítems que garanticen la equidad e interpretación fidedigna de las puntuaciones finales de la prueba, en correspondencia con los conocimientos reales adquiridos por el estudiante de satisfacerse esta demanda, la prueba quedaría sujeta a la pertinencia del contexto social, cultural y lingüístico del país y en consecuencia sería excelente.

4. Conclusiones

En el presente trabajo se propone una nueva taxonomía para la elaboración de pruebas de opción múltiple y se utilizó para analizar la prueba Planea para estudiantes de Educación Media Superior que se aplicó en 2017. Mostramos también que la taxonomía propuesta, se puede escribir como una matriz que sirve para evaluar cualquier tipo de test de opción múltiple y es, además, comprensible por el público versado en evaluación y conocedor del lenguaje de evaluación que utilizan los administrativos de la Secretaría de Educación Pública. Dicha matriz, dese luego, tiene validez general, pero constreñimos su aplicación a evaluar la prueba antes mencionada y encontramos que el 10% de sus ítems tiene errores de validez en su construcción, quedando muy lejos de una calificación de excelente. Además, se encontraron ítems que contravienen los lineamientos establecidos por CENEVAL para la elaboración de una prueba de opción múltiple. Este resultado tiene un impacto trascendente, pues es válido pensar que estos errores contribuyen al pobre posicionamiento de nuestros estudiantes en las pruebas que aplica la OCDE. Si este porcentaje impacta directamente sobre el puntaje obtenido, implica que el 6.62% de alumnos en el Nivel 1, cometió errores gracias a la redacción deficiente de los ítems aplicados por Planea EMS en 2017.

Agradecimientos
Vaya nuestro más sincero agradecimiento a Gustavo Hernández, Antonio Linares Flores y Sergio Vázquez Tlatoa por el apoyo computacional que nos brindaron.

5. Referencias.

Adam, P.& Caldwell D. J. (2014). Effects of multiple-choice item-writing guideline utilization on item and student performance. Recuperado de https://www.sciencedirect.com/science/article/pii/S1877129713001524 el 20 de mayo de 2017.

Arias, M. D. R. M. (1995). *Psicometría: Teoría de los tests psicológicos y educativos*. Síntesis. Recuperado de https://dialnet.unirioja.es/servlet/libro?codigo=82886 el 15 de mayo de 2017.Centro Nacional de Evaluación para la Educación Superior. (2016).

Diario Oficial. (2008). *Acuerdo Secretarial 444.* Recuperado de http://www.sems.gob.mx/work/models/sems/Resource/10905/1/images/Acuerdo_444_marco_curricular_comun_SNB el 10 de mayo de 2017.

Haladyna, T. M., Downing, S. M. y Rodríguez, M. C. (2002). *A review of multiple-choice item-writing guidelines for classroom assessment. Applied measurement in education*. 15(3), 309-333. Recuperado de https://scholar.google.com/scholar?q=Haladyna%2C+T.M.%2C+Downing%2C+S.M.+y+Rodr%C3%ADguez%2C+M.C.+%282002%29.+A+review+of+multiple-choice+item-writing+guidelines.+Applied+Measurement+in+Education%2C+15+%283%29%2C+309-334.&btnG=&hl=es&as_sdt=0%2C5 el 20 de mayo 2017.

Instituto Nacional para la Evaluación de la Educación. (2017). *Plan Nacional para la Evaluación de los Aprendizajes (Planea). Resultados nacionales 2017. Educación Media Superior. Lenguaje y Comunicación; Matemáticas.* Recuperado de http://planea.sep.gob.mx/content/general/docs/2017/ResultadosNacionalesPlaneaMS2017.PDF el 26 de octubre 2017. INEE.

Messick, (1989). American Psychologist, Vol. 35(11), Nov 1980, 1012-1027. Recuperado de http://psycnet.apa.org/index.cfm?fa=buy.optionToBuy&id=1981-27017-001 el 20 de mayo de 2017.

Moreno, R., Martínez, R. J., y Muñiz, J. (2004). Directrices para la construcción de ítems de elección múltiple. *Psicothema*, *16*(3), 490-497. Recuperado de http://www.redalyc.org/comocitar.oa?id=72716324 el 20 de Junio 2017.

Oliden, P. E. (2003). *Sobre la validez de los test*. *Psicothema*, *15*(2), 315-321. Recuperado de http://www.psicothema.es/pdf/1063.pdf el 15 de mayo de 2017.

Organización para la Cooperación y el Desarrollo Económicos. (2016). *PISA 2015 Resultados clave.* Recuperado de https://www.oecd.org/pisa/pisa-2015-results-in-focus-ESP.pdf el 15 de Mayo 2017. OCDE.

Osterlind, S. J. (1998). *What Is Constructing Test Items?* (pp. 1-16). Springer Netherlands. Recuperado de https://link.springer.com/chapter/10.1007/0-306-47535-9_1#page-1 el 20 de mayo de 2017.

Rodríguez, N. (1999). *Glosario de términos psicométricos y áreas afines.* Recuperado de http://www.psicoconsult.com/getattachment/ el 28 de mayo de 2017. a31bf150-b25a-4c13-b0b9-cf2dcb3a55e6/Glosario-de-Terminos.aspx el 20 de mayo de 2017.

Secretaría de Educación Pública. (2016). *Manual para el usuario Planea 2016*. Recuperado de http://planea.sep.gob.mx/content/ms/docs/2016/manuales/Manual_usuarios2016.pdf el 20el 22 Mayo de 2017. SEP.

Secretaría de Educación Pública. (2017a). *PLANEA 2017*. Recuperado de http://planea.sep.gob.mx/ms/aplicacion/documentos_normativos el 28 de agosto de 2017.SEP.

Secretaría de Educación Pública. (2017b). *Examen PLANEA 2017 en línea* Recuperado de http://planea.sep.gob.mx/ms/aplicacion/documentos_normativos el 20 de agosto de 2017. SEP.

Secretaría de Educación Pública. (2017c). *Modelo Educativo para la educación obligatoria*. Recuperado de https://www.gob.mx/cms/uploads/attachment/file/198738/Modelo_Educativo_para_la_Educacio_n_Obligatoria.pdf el 20 de agosto 2017. SEP.

Tristán. (2017). *La Objetividad en las Pruebas Estandarizadas.* Recuperado de https://revistas.uam.es/index.php/riee/article/viewFile/7592/7891 el 20 de mayo de 2017.