Se propone una estrategia para el análisis de datos que se recopilen a través de los instrumentos de medición aplicados en el marco de “las valoraciones relativas a los procesos de selección para la admisión y la promoción en educación básica y educación media superior en el marco del sistema para la carrera de las maestras y los maestros”. Al respecto, es importante distinguir entre dos tipos de instrumentos, para los cuales la aproximación analítica difiere en aspectos importantes: instrumentos de rendimiento óptimo e instrumentos de rendimiento típico. En este documento se describe la estrategia propuesta para los **instrumentos de rendimiento óptimo**. En el marco de las valoraciones relativas a los procesos de selección mencionados, estos instrumentos son de *opción múltiple* (con una opción correcta y varios distractores) y se asume que se habrán construido con base en un proceso que otorga la evidencia necesaria de validez para el uso (o los usos) que se pretenden dar a los resultados obtenidos por los mismos.

Por consiguiente, un primer paso importante del análisis de datos que se obtengan después de la administración de la prueba a la totalidad de participantes consiste en reconfirmar en esta muestra completa la evidencia de validez establecida en la fase de la construcción del instrumento. La fuente de validez relevante en este momento se basa en estudiar la estructura interna de la prueba (véase, para una introducción a las distintas fuentes de validez, los *Standards for Educational and Psychological Testing* publicados por American Educational Research Association, American Psychological Association, National Council on Measurement in Education, y Joint Committee on Standards for Educational and Psychological Testing, 2014). La estrategia de análisis que proponemos en este documento enfocará en tres tipos de análisis de la estructura interna: (1) un análisis de ítems en el marco de la Teoría Clásica de los Tests, (2) una evaluación de la bondad de ajuste del modelo psicométrico que subyace la prueba (el cual, para el caso de las pruebas de rendimiento óptimo, se sitúa en el marco de los modelos de diagnóstico cognitivo, véase la discusión más adelante) y (3) la aplicación de un procedimiento para detectar posible funcionamiento diferencial de los reactivos que componen la prueba. Las Secciones 1 a 3 de este documento proveerán más detalles de estos análisis, respectivamente.

Después de evaluar y realizar posibles ajustes para reconfirmar la evidencia de validez, se explicará cómo se deriva la calificación de cada sustentante en la prueba y cómo esta calificación contribuye a la toma de decisiones (específicamente, la lista de prelación; Sección 4). A pesar de que la selección (y ordenamiento) de los aspirantes con base en los resultados constituye el uso principal de la prueba—y la validación se centra en este uso—, se consideran dos objetivos secundarios de las pruebas de rendimiento óptimo: (1) proveer retroalimentación individual a los aspirantes sobre las habilidades que requieren apoyo y (2) presentar un reporte con un resumen de los resultados desagregados a nivel de las entidades federativas con el fin de formular recomendaciones para la política pública y/o las escuelas normales sobre las áreas de oportunidad en las habilidades valoradas en los aspirantes. La Sección 5 describe los análisis que se llevarán a cabo para lograr estos objetivos secundarios.

**1.  Análisis de ítems en el marco de la Teoría Clásica de los Tests**

En esta sección se describirán (1a) los distintos índices, de dificultad y discriminación, que vamos a calcular (es decir, su definición y/o la forma de calcularlos), (1b) cómo se interpretan estos índices (es decir, su significado), (2a) el método gráfico para un análisis visual del ítem, incluyendo los distractores, (2b) cómo interpretar esta gráfica y (3) la forma (los criterios y procedimientos) de tomar decisiones con base en estos resultados (mencionando aquí la importancia del contenido de los ítems y por lo tanto el consultar a los equipos involucrados en el desarrollo de la prueba).

**2.  Bondad de ajuste de un modelo de diagnóstico cognitivo**

Los instrumentos de rendimiento óptimo se construirán en el marco de los modelos de diagnóstico cognitivo (véase, por ejemplo, de la Torre y Minchen, 2014); estos modelos se pueden considerar como modelos multidimensionales de la Teoría de Respuesta al Ítem (TRI).

Aquí describimos el procedimiento para evaluar la bondad de ajuste del MDC, cuáles son los pasos, en cada paso las decisiones que se tomarán y los criterios para las mismas, qué hacer cuando se detecta discrepancias significativas entre el modelo y los datos, etc. Puede ser relevante el capítulo de Han y Johnson (2019) que habla de los índices para evaluar el ajuste global y a nivel de los ítems en los MDCs.

**3.  Funcionamiento diferencial del ítem**

En esta sección, primero se define DIF (con un ejemplo). Posteriormente presentamos los métodos que permiten detectar DIF (aquí tendremos que buscar cuáles son las estrategias más adecuadas; puede ser útil el capítulo de Qiu, Li, y Wang, 2019). También vamos a describir cómo lidiar con DIF en caso de que se detecte en los instrumentos que estamos analizando.

**4.  Calificaciones globales en la prueba**

En esta sección se elaborarán dos elementos. Por un lado, nuestro equipo es responsable de calcular las calificaciones; por lo tanto, explicaremos cómo se derivan estas calificaciones a partir de las respuestas de cada sustentante en la prueba y cuáles son los controles de calidad que se realizan. (Opino que es muy importante, dado el alto impacto de las pruebas, reflexionar sobre cómo asegurar que no hay errores en los cálculos.) La fórmula exacta, que explicita cómo integrar las calificaciones de los distintos componentes evaluados por el MDC, se acordará en consulta con el equipo de diseño de pruebas. Por otro lado, estas calificaciones se juntarán con los resultados en las otras pruebas incluidas en el proceso de selección para llegar a un resultado final (una lista de prelación). Aunque esta integración es responsabilidad final de la SEP-USICAMM, nosotros les daremos, de modo de ilustración, unos ejemplos (con base en resultados ficticios para las otras pruebas incluidas en el proceso de selección) de cómo llevarla a cabo de una manera correcta y segura.

**5.  Resultados por componente evaluado para retroalimentación individual y política pública**

**5.1.  Resultados para generar los reportes individuales**

Como hemos hablado anteriormente, se generarán reportes individuales sobre el desempeño de cada aspirante, indicando cuáles son las áreas de oportunidad o que deben ser atendidas para generar un mejor efecto en los educandos. Nuestro equipo proporcionará los resultados detallados por sustentante al equipo de sistemas. En esta sección se explica brevemente cómo se obtienen dichos resultados.

**5.2.  Reportes con resultados a nivel nacional y por entidad federativa**

Además de los reportes individuales, la SEP-USICAMM pide un reporte en el cual se presentan—en tablas y gráficas—e interpretan los resultados resumidos a nivel nacional (tanto la calificación global en la prueba, como los resultados por componente evaluado en al MDC). Estos reportes pueden orientar a la política pública para desarrollar o proponer iniciativas que mejoren la preparación de las maestras y maestros. También piden presentar (en un anexo) las tablas y gráficas de los mismos resultados, pero desagregados por entidad federativa. En esta sección se describe cómo se realizan los análisis para generar estos reportes y daremos un ejemplo para ilustrar el diseño/organización de dichos reportes.

**Referencias**

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, y Joint Committee on Standards for Educational and Psychological Testing (U.S.). (2014). *Standards for educational and psychological testing.* Washington, DC: AERA.

de la Torre, J., y Minchen, N. (2014). Cognitively diagnostic assessments and the cognitive diagnosis model framework. *Psicología Educativa*, *20*(2), 89–97. doi: [10.1016/j.pse.2014.11.001](http://dx.doi.org/10.1016/j.pse.2014.11.001)

Han, Z., y Johnson, M. E. (2019). Global- and item-level model fit indices. En M. von Davier y Y.-S. Lee (Eds.), *Handbook of diagnostic classification models: Models and model extensions, applications, software packages* (pp. 265–285). Cham, Suiza: Springer. doi: [10.1007/978-3-030-05584-4\_13](https://doi.org/10.1007/978-3-030-05584-4_13)

Qiu, X.-L., Li, X., y Wang, W. C. (2019). Differential item functioning in diagnostic classification models. En M. von Davier y Y.-S. Lee (Eds.), *Handbook of diagnostic classification models: Models and model extensions, applications, software packages* (pp. 379–394). Cham, Suiza: Springer. doi: [10.1007/978-3-030-05584-4\_18](https://doi.org/10.1007/978-3-030-05584-4_18)