REVIEWER A

Primera revisión

El artículo está, en general, bien escrito y resume un área de investigación interesante. Señalaré los aspectos que creo que pueden necesitar trabajo adicional. Empiezo por los aspectos más relevantes desde mi punto de vista:

- Sorprende un artículo en el que la introducción llega hasta la página 21 y le siguen únicamente cinco páginas de resultados. Uno de los objetivos manifestados es ser un artículo ilustrativo. Creo que se pierde parte de la eficacia en este punto con la premura en la presentación de resultados. Por ejemplo, una tabla con las diferentes estimaciones de los parámetros por ítem para los dos modos de calibración haría más sencillo el ir guiando al lector a lo largo del proceso de extracción del resto de resultados. De otro modo, el artículo corre el riesgo de caer en tierra de nadie, entre aquello que puede necesitar una persona con conocimientos medios-avanzados de psicometría y aquella persona con conocimientos básicos.
- Con respecto a la relación entre el modelo lineal y el GRM, algunos de los resultados presentados por los autores han sido descritos en publicaciones previas. Por ejemplo, la relación en S entre las puntuaciones estimadas por ambos modelos se encuentra en Dumenci y Achenbach, 2008.
- Los autores discuten el índice l_z como si éste estuviera libre de limitaciones y como si no se hubieran propuesta mejores del mismo. Véase, por ejemplo, Magis, Raîche, y Béland, 2012.
- En la ecuación 15 de evaluación del cambio se asume que la medida retest carece de error. Esto no es defendible. Por ello, sería necesario utlizar indicadores de cambio que incluyen estimaciones de la precisión de medida para ambas ocasiones. Un punto de lectura que podría ser de utilidad para esto sería Finkelman, Weiss, y Kim-Kang (2010).
- A lo largo de todo el texto los autores no se ciñen al estilo de la APA. Por ejemplo, y sin afán de exhaustividad, el texto no ha de ir justificado, sino alineado a la derecha, segundas citas o posteriores para Pallero, Ferrando, & Lorenzo-Seva (2006) tendrían que ser Pallero et al. (2006) –igual aplica para otros artículos–, los niveles de texto (‘headings’) no siguen la estructura prescrita, series de citas dentro de paréntesis en ocasiones se separan con comas cuando corresponde hacerlo con puntos y coma… Considero que los autores tendrían que realizar un esfuerzo adicional para ceñirse al estilo adoptado por la revista.

Igualmente, creo que hay otros aspectos menores que también necesitan mejorarse:

- En la presentación de resultados, se indican como líneas de referencia dos desviaciones típicas separadas de las medias para los dos índices. No se ofrece justificación para este valor.

- Si el objetivo es evaluar si los dos indicadores llegan a marcan como ajustados/desajustados a los mismos examinados, podría ser adecuado emplear una medida de acuerdo para variables categóricas.
- Entiendo que la figura 2b apenas aporta, ya que es equivalente a la figura 2a con otra escala en el eje de las Y, puesto que es el resultado de dividir por una constante. Sería más sencillo el incluir esta doble información en una sola figura.
- En la introducción se comenta que el test CTAC presenta características adecuadas para ser tratado con el modelo lineal. Sería oportuno introducir los argumentos numéricos que apoyan este planteamiento en este punto del artículo.
- En la página 5 se manifiesta que ahora se cuenta con una muestra representativa. ¿En base a qué criterios es esta muestra representativa?
- La descripción del GRM creo que tendría que hacerse no para el caso específico del CTAC (cinco alternativas de respuesta), sino en su formulación genérica. Tendrían que hacerse los cambios oportunos en las fórmulas para ello.
- Sería oportuna una referencia para la ecuacion 8.
- Los autores han de confirmar que todos los símbolos empleados en las ecuaciones son correctados definidos en el texto, y en posiciones cercanas a las primeras ecuaciones donde aparecen tales símbolos. Por ejemplo, k, n, ω…
- ¿Cuál es el significado de esta frase "The data were collected and introduced by the Tarrragona psychologist"?
- Sería conveniente que los autores incluyeran los puntos de corte para los índices de ajuste que van a emplear para marcar un modelo como con ajuste aceptable. Según algunos de estos puntos de corte, por ejemplo, se consideraría que el modelo lineal presenta un ajuste insatisfactorio.
- Los autores indican que la máxima información se consigue para el rango de niveles de rasgo donde se encuentra la mayor parte de los evaluados. Creo que la redacción de esta parte puede dar lugar a confusiones. Este resultado se debe a los valores de los parámetros de localización de los ítems, cuya media muy posiblemente coincida con la media de θ. Pero no siempre ha de ser así.
- Información sobre la curtosis de los ítems sería oportuna.
- No he conseguido comprender con claridad por qué los autores consideran que los examinados con puntuación extrema (que no "extreme respondents") son mejor evaluados por el modelo lineal.
- De hecho, el que patrones de respuesta constantes y extremos (todo 0 o todo 4) no se correspondan con un nivel de rasgo estimado de ±infinito con estimación máximo-verosímil no es algo intuitivo, salvo que se ponga especial atención a la lógica del modelo lineal y a la fórmula de las puntuaciones factoriales de Barlett.
- Otros artículos que guardan relación con los temas tratados por los autores serían Kamata y Bauer (2008) y Maydeu-Olivares, Cai, y Hernández (2011).

Referencias

Dumenci, L., & Achenbach, T. M. (2008). Effects of estimation methods on making trait-level inferences from ordered categorical items for assessing psychopathology. *Psychological Assessment, 20*, 55–62.

Finkelman, M. D., Weiss, D. J., & Kim-Kang, G. (2010). Item selection and hypothesis testing for the adaptive measurement of change. *Applied Psychological Measurement, 34*, 238–254.

Kamata, A, & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling, 15*, 136–153.

Magis, D., Raîche, G., & Béland, S. (2012). A didactic presentation of Snijders's l_z* index of person fit with emphasis on response model selection and ability estimation. *Journal of Educational and Behavioral Statistics, 37*, 57–81.

Maydeu-Olivares, A., Cai, L., & Hernández, A. (2011). Comparing the fit of item response theory and factor analysis models. *Structural Equation Modeling, 18*, 333–356.

Segunda revisión

En primer lugar, felicitar a los autores por los cambios introducidos en el artículo. Considero que ahora los objetivos son más adecuados y quedan mejor cubiertos. Pase a describir algunas cuestiones que creo que necesitan de algo más de trabajo y alguna clarificación adicional.

- Uno de los temas recurrentes en el artículo es la relación en forma de S entre las estimaciones del nivel de rasgo por parte del GRM y el estimaciones del modelo lineal. No he conseguido entender con claridad cómo justifican los autores esta relación. No digo que no sea la esperable, sino que creo no haber visto reflejados en el artículo argumentos claros para su defensa.

- En la misma línea, creo que no se argumenta bien las razones para una mayor dispersión de las respuestas con el GRM en el caso de patrones de respuesta extremos. Con estimación ML, un patrón de respuestas constante y extremo (todo puntuación mínima o todo puntuación máxima) lleva a una estimación de theta igual a infinito (negativo o positivo). Sin embargo, con el modelo lineal una sólo una respuesta igual a infinito (fuera de rango e imposible) supone una estimación igual a infinito. Por ello, el modelo lineal conlleva una 'regresión a la media' similar a la esperable con estimación bayesiana del nivel de rasgo (tipo de estimación nada infrecuente en TRI).

- En la nueva versión del artículo el foco pasa a estar principalmente en la escala CTAC. Por ello, creo que sería oportuno mostrar información, ítem a ítem, de algunos de los resultados obtenidos: umbrales, cargas, interceptos... De otro modo, creo que sigue dando la impresión de que la escala CTAC es una 'excusa' para el artículo y lo relevante es únicamente la parte de comparación entre aproximaciones de análisis.

- La solución común para establecer la conveniencia de una puntuación general para escalas multidimensionales son los modelos bifactoriales. Entiendo que ajustar un modelo bifactorial queda alejado de los objetivos del artículo, pero creo que sería adecuado comentar las aproximaciones más adecuadas para algunos de los problemas que se plantean con la escala.

- Quizá estoy equivocado en este punto, pero diría que los ji-cuadrados de los dos modelos ajustados no son directamente comparables, al provenir de estimaciones MLR y WLSMV.

- No se ofrece ninguna referencia para la relación entre discriminación de los ítems e índices de ajuste (pág. 16).

- ¿Cómo es posible que al emplear la Ecuacion 4, que tiene en cuenta dos fuentes de incertidumbre (estimación pre y estimación post) el intervalo de confianza sea más estrecho que cuando se tiene en cuenta una sola fuente de información?

- En la página 19, cuando se habla de detección, se está hablando de una variable dicotómica, luego entiendo que no tiene especial sentido el comentar la fiabilidad o el error típico, sino que es más informativo el porcentaje de clasificaciones correctas o algún indicador de corrección en clasificaciones binarias.

- Una parte de los argumentos que se ofrecen para preferir el modelo lineal por encima del GRM en la discusión podría llevar a preferir la teoría clásica de los tests sobre el modelo lineal. Creo que sería oportuno afinar algo más este apartado.

- En ocasiones se escriben frases que pueden lugar a confusión, como "el GRM mide con mayor precisión en este rango" (p. ej., página 18). Hay que dejar claro que, en todo momento, se está hablando de estimaciones, no de la realidad. Porque, si no, la conclusión sería: si se desea evaluar con mayor precisión en nivel extremos, emplee el modelo lineal, dado que así es mayor la información obtenida. Y, claramente, esto es incorrecto.

- Convendría aclarar si hay un cierto solapamiento entre la muestra empleada en este estudio y la empleada en estudios previos.

- En la Ecuación 6 quedan por definir varios de los elementos que ahí aparecen.

- La cita de Ferrando et al. (2010) es algo forzada, puesto que entiendo que el que ese artículo sea hecho en España o publicado en España no es un valor especial para la psicometría.

- Se dice (p. ej., página 5) que el CTAC está pensado para detectar a personas con niveles algo de ansiedad. Sorprende que se hable de niveles alto dado que el punto de corte es una theta superior a 0.40, un valor más bien medio dentro de la población de referencia.

- La tesis de Kef se incluye entre las referencias de instrumentos en inglés. Sorprende esto, puesto que una tesis defendida en la Universidad de Amsterdam con muestra holandesa.

- No se ha incluido la referencia a Welsh (1997).

- Se habla de que la evaluación se hace un contexto de ansiedad severa (pág. 3). ¿La ansiedad es contextual (más bien, por ello, ansiedad-estado) o es ansiedad-rasgo?

- Cuando se elige entre el ítem 9 y el 27, no se aclara la razón para retener uno de los dos en concreto.

- El texto necesita ser revisado para acomodarse al manual de estilo de la APA. Presento algunos ejemplos. Las teorías o modelos no van con mayúsculas (no es Teoría de Respuesta el Ítem, sino teoría de respuesta al ítem). No pueden ir dos paréntesis encadenados. Cuando los autores son seis o más, se cita dentro del texto con Primer Autor et al. Cuando dos autores diferentes compartes apellido (p. ej., B. Muthén y L. K. Muthén) se incluyen las iniciales cuando se citan sus trabajos.

- El resumen en inglés ha de ser revisado. Especialmente, la frase que empieza por “Second”.

REVIEWER B

Cuando acepté la revisión de este artículo pensé en que se trataba de un estudio empírico donde se trataba de llevar a cabo una comparación entre modelos lineales y no lineales de TRI para evaluar las propiedades métricas de un test de ansiedad en deficientes visuales (en términos de precisión en la medida, sensibilidad al cambio y ajuste de las personas), tal como reza en su título y su resumen. Pero la lectura detallada me ha llevado a varias consideraciones al respecto.

Primero, creo que no estamos estrictamente ante un artículo empírico, sino que el texto tiene un desarrollo teórico-metodológico considerable, en los que los apartados de “The Linear Model and the Graded Response Model: Some Basic Comparisons”, de “Measurement Precision and Sensitivity to Change” y “Person-Fit Assessment” (especialmente el primero) son demasiado extensos en su desarrollo (van desde la página 8 a la 22). La información que se expone en ese primer apartado, que intenta mostrar cómo el modelo de respuesta graduada (TRI) y el análisis factorial (AF) pueden confluir en una formulación general del AF que posibilite la comparación entre aproximaciones lineales y no lineales, está tratado de más. Hasta la página 7 de la introducción la línea del artículo es adecuado, se revisa la variable objeto de medida y las formas y modelos utilizados para su medida. No sé si sería mejor hacer un artículo teórico-metodológico sobre el tema que insertarlo en un artículo como éste. Los mismos autores comentan que las comparaciones entre aproximaciones lineales y no lineales han sido frecuentes en la literatura (e.g. Ferrando, 1999; McDonald, 1982, 1999; Xu, & Stone, 2012), pero de una forma estrictamente teórica. Pues bien, lo que hacen en el apartado “The Linear Model and the Graded Response Model: Some Basic Comparisons” vuelve a ser un intento teórico de acercamiento entre ambas aproximaciones, cuestión que quizás está tratada demasiada extensamente en el artículo, y quizás con unas breves y resumidas pinceladas al respecto hubiese bastado. Pero, obviamente, esta es mi impresión. Algo parecido pero de menor magnitud ocurre con los dos apartados siguientes.

Una cuestión de fondo que me planteo es que si este artículo no trata de ver la ganancia (en términos de precisión en la medida, sensibilidad al cambio y ajuste de las personas) de aplicar un modelo lineal o no lineal para estimar las propiedades de la medida de ansiedad concreta (test CTAC), sino mas bien de ver qué modelo es mejor para estimar en términos de precisión en la medida, sensibilidad al cambio y ajuste de las personas en general, como parece desprenderse en algunos párrafos de las conclusiones (p.e. párrafo 3 página 29) y del desarrollo teórico de la introducción. Creo que en este caso hubiese sido mucho más oportuno haber planteado un estudio de simulación. Máxime cuando lo sustantivo del desarrollo del artículo deja en un segundo plano las propiedades del CTAC bajo ambos modelos.

Otra cuestión importante es que se comenta en el texto que el CTAC es "essentially a unidimensional instrument". Pero existe un artículo publicado en la revista Psicothema:

> Ferrando P.J, Lorenzo-Seva U. y Pallero R. (2009). Implementación de procedimientos gráficos y analíticos para la construcción de formas paralelas *Psicothema, Vol. 21, nº 2,* pp. 321-325)

donde se hace referencia a ese mismo CTAC (página 323), diciendo lo siguiente:

"El Cuestionario Tarragona de Ansiedad para Ciegos (CTAC; Pallero, Ferrando y Lorenzo, 2006) es un instrumento bidimensional, utilizado sobre todo en adultos que han perdido total o parcialmente la visión. El test tiene 35 reactivos en formato Likert de 5 puntos, y sus dos subescalas: Ansiedad Cognoscitiva (AC) y Ansiedad Fisiológica muestran propiedades psicométricas más que aceptables".

¿Cómo explican los autores del artículo en revision estos datos aparentemente contradictorios?

También el último párrafo de la página 7 del texto en revisión dice "Consider the set of 34 CTAC items, with a 5-point response format, that measure a single trait of anxiety θ". El CTAT tiene ¿35 o 34 ítems?

Por cierto, los autores del artículo en revision quizás deberían citar el artículo de Psicothema arriba indicado.

Entrando en cuestiones más concretas (y de menor importancia), quiero hacer las siguientes consideraciones:

1. En el apartado de análisis preliminares, y observando el rango de valores de la discriminación de los ítems, me pregunto por qué en vez de utilizar el GRM no se han utilizado modelos más parsimoniosos que entienden la existencia de una discriminación similar en los ítems. Estoy pensando en modelos de la familia de Rasch (p.e. Rating Scale Model o incluso el Partial Credit Model).
2. El concepto de evaluación que aparece en el título (Assesing) tiene una conceptualización mucho más amplia que testing o medición. Sin entrar en consideraciones al respecto, recomendaría cambiar la palabra en el título por medición. Incluso se podría quedar la palabra evaluación si va referido a las propiedades psicométricas consideradas, como así se dice en el primer objetivo del resumen.
3. Dado que la referencia Pallero, Ferrando, & Lorenzo-Seva (2006) a la que hacen referencia los autores no es una referencia susceptible de consulta en revistas especializadas, sería de gran interés para los lectores que los autores explicaran brevemente la definición operativa del constructo asi como que realizaran una breve descripción de los contenidos del test CTAC.
4. La siglas CTAC (Tarragona Anxiety Questionnaire for the Blind) no se corresponde con el nombre en inglés.

Algunas erratas a corregir:

- Página 5, Segundo párrafo (Pallero, Ferrando, & Lorenzo-Seva, 20006), debería ser "2006".

- La referencia de

  Pallero, R., Ferrando, P.J., & Lorenzo-Seva, U. (2006) “Cuestionario Tarragona de Ansiedad para Ciegos” Madrid: ONCE.

  debería ponerse en formato APA.
- Página 11, en la referencia (Samejima, 1969 Baker, 1992) habría que poner una coma para separar ambas referencias.

- La referencia (Olsson, 1979, Muthén & Kaplan, 1985) de la página 12 debería ponerse en orden alfabético.
- En la ecuación (4) debería especificarse que $\tau_{j.}$ es el parámetro de umbral de respuesta.