REVIEWER A

The manuscript compares three approaches for detecting polytomous Differential Item functioning. The paper is well written and appropriate for *Psicologica* in relation to its topic, but I have some concerns:

1) Method (v.g., MACS, IRT) is confounded with strategy (vg., free baseline model vs constrained baseline model). As expected, IRT-LR and logistic regression type I error rates are inflated in some cases because the anchor is contaminated and no purification procedure is applied. This is not new and some purification method should be applied.

2) Furthermore, for MACS it is assumed that a DIF-free anchor item is known. This is not realistic. If the free-baseline strategy is used, they should pay more attention to the procedures for detecting the free-DIF anchor items (Woods, 2009).

3) Only one test length is simulated, only absence of impact is simulated, for MACS, only one anchor-referent item is studied (15th item),... All of these limit generilizability of results. More conditions should be simulated.

4) Previous studies with the ordinal logistic regression are not described. They should be described.

Minor aspects:

1) Linear CFA may not be optimal analysis for ordered-categorical data. Measurement invariance testing under ordinary linear CFA is not adequately comparable to differential item functioning (DIF) analysis under IRT. Ordinal methods can be applied for MACS and this comparison may be performed (see Kim and Yoon, 2012). Some comment to this is required.

2) No correctionis applied for multiple comparisons.Why?

3) The methods presentation is not balanced. Advantages of observed scores methods are underlined but disadvantages are not described.

References:

Kim, E. & Yoon, M. (2011): Testing Measurement Invariance: A Comparison of Multiple- Group Categorical CFA and IRT, Structural Equation Modeling: A Multidisciplinary Journal, 18, 212-228

Woods, C. M. (2009). Empirical selection of anchors for tests of differential item functioning, Applied Psychological Measurement, 33, 42-57.

REVIEWER B

El artículo está escrito con gran claridad y responde a los objetivos planteados. El estudio de simulación que constituye el grueso del artículo es relevante y parece bien ejecutado. Los cambios que se sugieren son, por tanto, menores aunque pertinentes.

1.- En la introducción se señala: “Although there are some interesting papers about the detection of DIF in polytomous data topic (French & Miller, 1996;

Kristatkansson, Aylesworth, McDowell, & Zumbo, 2005; Spray & Miller, 1994; Zwick, Donogue & Grima, 1993), the level of development regarding the detection procedures for polytomous items has not been widely studied." Hace 10 años, la anterior frase sería totalmente cierta, hoy en día no es correcta. Se deben citar los artículos más relevantes que sobre la detección del DIF en ítems politómicos se han publicado en los últimos 10 años. Aconsejo una breve revisión bibliográfica sobre el particular.

2.- En el apartado *DIF detection methods* se señala: "Observed-score based procedures focus on examining the relationship between item performance and observed scores, often based on total scores. Examples of observed-score based procedures are the Mantel-Haenszel procedure (Mantel, 1963; Potenza & Dorans, 1995). Se deberían citar los artículos más relevantes que en relación con ítems politómicos hay en cada técnica. Por poner un ejemplo, en el caso de los métodos MH se cita el trabajo de Mantel del 63 en el que se propone el test de Mantel, pero deberían citarse además: a) el artículo de Mantel & Haenszel (1953) en el que propusieron el estadístico MH generalizado, b) el artículo de Zwick, Donoghue & Grima (1993) donde se plantea por primera vez el uso de dichos estadísticos para detectar el DIF en ítems politómicos, y finalmente c) los artículos de Fidalgo & Madeira (2008) y Fidalgo & Scalon (2010) donde se aplica el estadístico MH generalizado formulado por Landis, Heyman & Koch (1978) en la investigación del DIF en ítems politómicos. Consideraciones análogas cabe hacer sobre los otros dos procedimientos. De nuevo, aconsejo una breve revisión bibliográfica sobre el particular.

3.- En esa misma página hay un pequeño error de concordancia "The advantages of observed-score methods is" en lugar de "The advantages of observed-score methods are"

4.- En la detección del DIF con la RL se utiliza un test con 2 grados de libertad que permite detectar tanto el DIF uniforme como el uniforme. Sería conveniente discutir la posibilidad, y las consecuencias, de ajustar modelos diferentes para detectar el DIF uniforme y el nouniforme, ya que cada uno de ellos tendría 1 gl y, por tanto, mayor potencia.

5.- Finalmente, en la leyenda de las tablas con el error de tipo I, sería conveniente señalar el nivel de significación utilizado en los análisis.

Espero que estos comentarios sirvan para mejorar un artículo ya de por sí excelente.