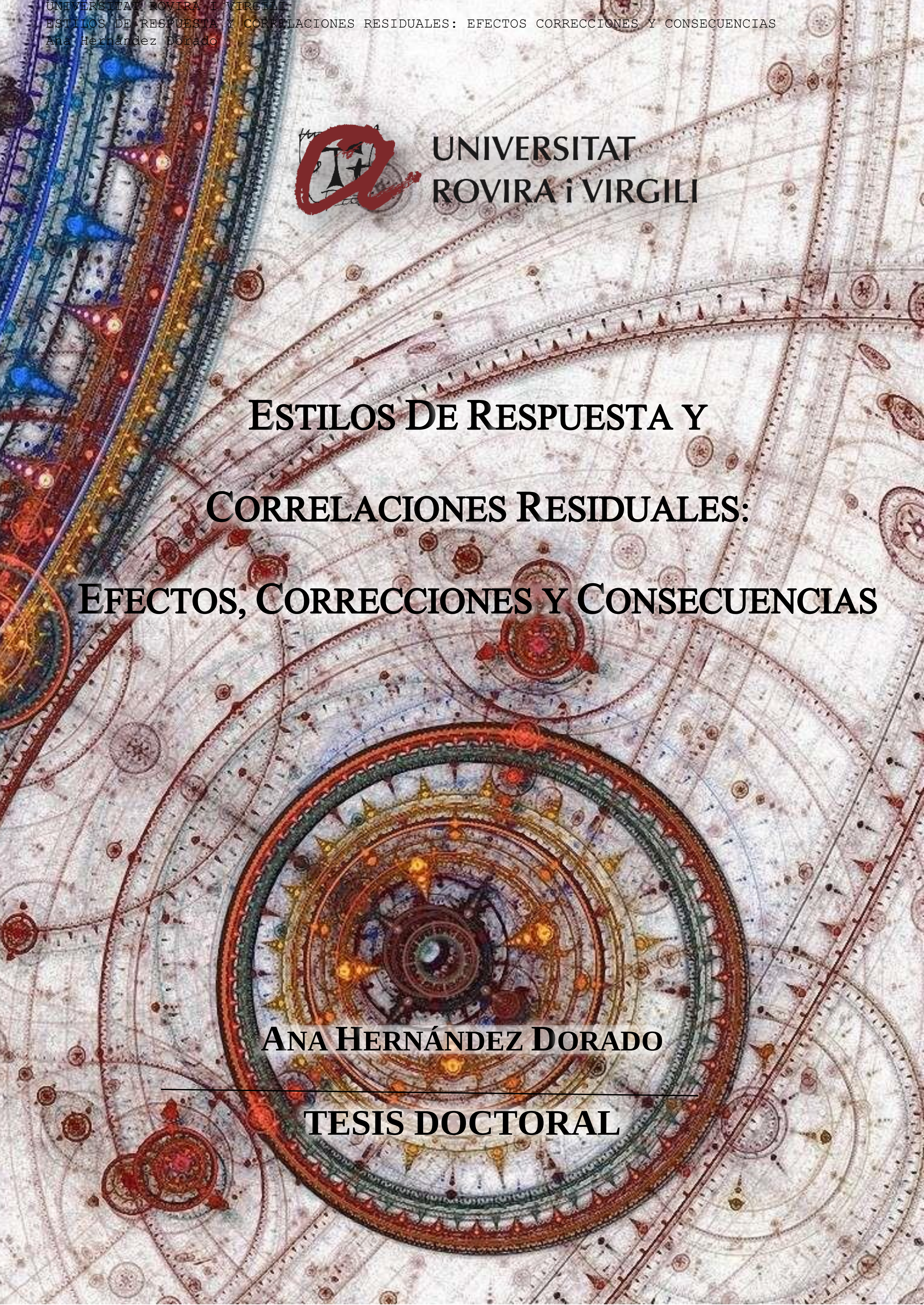
UNIVERSITAT ROVIRA I VIRGILI

ESTILOS DE RESPUESTA Y CORRELACIONES RESIDUALES: EFECTOS CORRECCIONES Y CONSECUENCIAS

Ana Hernández Dorado



UNIVERSITAT ROVIRA i VIRGILI

ESTILOS DE RESPUESTA Y CORRELACIONES RESIDUALES: EFECTOS, CORRECCIONES Y CONSECUENCIAS

ANA HERNÁNDEZ DORADO

TESIS DOCTORAL

Ana Hernández Dorado

ESTILOS DE RESPUESTA Y CORRELACIONES RESIDUALES: EFECTOS, CORRECCIONES Y CONSECUENCIAS

TESIS DOCTORAL

DIRIGIDA POR:

Dr. Pere Joan Ferrando

Dr. Andreu Vigil Colet

Departamento de Psicología

UNIVERSITAT
ROVIRA i VIRGILI

Tarragona

2024

… la parte contratante de la primera parte será considerada como la parte contratante de la primera parte…

…la parte contratante de la segunda parte será considerada como la parte contratante de la segunda parte…

… ¿por qué no hacemos que la primera parte de la segunda parte contratante sea la segunda parte de la primera parte?

UNA NOCHE EN LA ÓPERA

FAIG CONSTAR que aquest treball, titulat "Estilos de Respuesta y Correlaciones Residuales: Efectos, Correcciones y Consecuencias", que presenta Ana Hernández Dorado per a l'obtenció del títol de Doctor, ha estat realitzat sota la meva direcció al Departament de Psicologia d'aquesta universitat.

HAGO CONSTAR que el presente trabajo, titulado "Estilos de Respuesta y Correlaciones Residuales: Efectos, Correcciones y Consecuencias", que presenta Ana Hernández Dorado para la obtención del título de Doctor, ha sido realizado bajo mi dirección en el Departamento de Psicología de esta universidad.

I STATE that the present study, entitled "Estilos de Respuesta y Correlaciones Residuales: Efectos, Correcciones y Consecuencias", presented by Ana Hernández Dorado for the award of the degree of Doctor, has been carried out under my supervision at the Department of Psychology of this university.

Tarragona, 23 de novembre de 2023

El/s director/s de la tesi doctoral
El/los director/es de la tesis doctoral
Doctoral Thesis Supervisor/s

Pere Joan Ferrando

Andreu Vigil Colet

Agradecimientos

En primer lugar, quiero agradecer a Andreu y Pere Joan no sólo por aceptar dirigir esta Tesis, sino por la infinita paciencia y ayuda prestada, tanto a nivel profesional como personal, para que esta llegara a buen término. Porque en palabras del Dr Vigil y parafraseando a Hesíodo: “al principio era el Caos, y del Caos surgió Ferrando”. Nunca una broma fue tan cierta. Habéis sido tenaces en vuestra entrega y dedicación en estos años.

También mi gratitud al Dr. Urbano Lorenzo por ser una fuente de conocimiento constante, por hacer de tutor sin serlo, por tu sabiduría y tu interés en mejorar mi carrera investigadora. A la Dr. Fàbia Morales por tu apoyo y confianza al permitir colaborar en tus proyectos; eres excelente, tanto a nivel profesional como personal, y considero un placer trabajar contigo. Por supuesto, mi total agradecimiento a todos los compañeros del IPSII. Siempre estará en mi corazón todos los buenos momentos que pasé en Córdoba. Fue una experiencia realmente enriquecedora tanto a nivel personal como profesional.

A mis compañeros que han estado en los descansos, mi más sincera gratitud por hacerme sentir como en casa, arropada y darme ese punto de humor necesario para aguantar todo el día: Ivette, Jordi, Ana, Pi, Marcos, Joan, Cinta, Cristina, Esther, Estrella…Un agradecimiento especial a Antoni Masip por tu compañía diaria, ayuda y soporte informático y a Esther Carrasco por la disposición y paciencia en cada duda con el “papeleo del demonio”; ambos habéis estado acompañándome y guiándome en los momentos de pánico.

A Laura, Víctor y Marina porque desde que os conozco escucháis mis aventuras, desventuras y vicisitudes. Por último, pero no menos importante, a mi familia: Alberto, María Rosa y Lidia por estar siempre que lo necesito y por ser el pilar de mi vida. Aún en la distancia sois los que más habéis sufrido mis accidentes, caídas y altibajos.

Índice

RESUMEN. ORGANIZACIÓN DE LA TESIS

El presente documento está distribuido en cinco capítulos a partir de los cuales se expone el tema de los sesgos de aquiescencia y dependencia local. A lo largo del Trabajo se expondrá el marco teórico, la metodología utilizada a lo largo de los artículos publicados, los resultados y las conclusiones a las que se ha llegado.

El primer capítulo, intentará dar una visión global del estado de la cuestión. Cómo ha ido evolucionando el estudio de la aquiescencia y la dependencia local, su importancia, las diferentes líneas de investigación y los modelos más significativos.

El segundo capítulo se centra en los objetivos que se plantean en esta investigación y las hipótesis relacionadas con los mismos.

El tercer capítulo hablará de la metodología utilizada en los artículos que componen la tesis. Se subdividirá de forma que cada sección explique el procedimiento que se ha llevado a cabo. Se describirán los instrumentos utilizados en el estudio y el tipo de variables que se evalúan.

El cuarto se centra en los resultados obtenidos mostrando los artículos publicados, aceptados o en revisión.

El último capítulo comentará de forma global los resultados anteriores considerando las implicaciones teóricas de los mismos. Además, se especificarán las limitaciones del presente estudio y los nuevos interrogantes que se deben plantear de cara a una futura investigación

Capítulo 1. Introducción

De sobra nos es conocido que uno de los mayores retos en el campo de la psicología es la medición de los rasgos de personalidad. En las últimas décadas ha habido un gran auge en la creación y adaptación de test psicológicos, que son, en su mayor parte, autoinformes. Los autoinformes, a pesar de ser una herramienta valiosa para recopilar información sobre el rendimiento típico de los individuos, presentan algunas limitaciones, siendo los sesgos de respuesta, sesgos de memoria, sesgos culturales, y limitaciones de la autoevaluación las principales. Cuando se presenta alguno (o varios) de estos problemas, la principal consecuencia es que las respuestas a los ítems no tienen una interpretación unívoca al estar influidas por otros determinantes aparte del contenido que se pretende medir. El presente trabajo se enfocará en el impacto de dos de los potenciales problemas: la aquiescencia y la independencia local, tanto por separado como en su conjunto.

Tradicionalmente la aquiescencia (ACQ) se define como la tendencia de los participantes a estar de acuerdo a los ítems independientemente del contenido de estos (Paulhus and Vazire, 2005). La ACQ es uno de los estilos de respuesta más estudiados: como único distorsionador (Bentler et al., 1971) o acompañado por otros sesgos de respuesta como la deseabilidad social (DS) (Ferrando et al., 2009; Navarro-Gonzalez et al., 2016). Por otro lado, podemos definir a los residuales correlacionados o, más en general, al problema de la dependencia local como el fenómeno por el cual los ítems covarían de forma sistemática sin que este hecho se pueda explicar únicamente por la variable subyacente (Embretson y Reise, 2013).

1. Estudio de la Aquiescencia

Ahora bien, ¿por qué se produce la aquiescencia? ¿Es un rasgo de personalidad inherente al ser humano? ¿Un problema debido al contexto o a la forma de administrar los ítems? ¿Se debe al diseño o a la propia redacción de los ítems?

Pueden distinguirse en la literatura dos vertientes generales en la investigación: la que enfoca la aquiescencia desde las características de los ítems que, supuestamente, la elicitan, y la que se centra en las características de los individuos (mayor o menor predisposición a dar este tipo de respuesta). ¿Cómo se explicaría pues la ACQ?

1.1. *Metateorías de la Aquiescencia: Teorías motivacionales vs cognitivas*

En la literatura, existen varios modelos que explicarían la aquiescencia como debida a un proceso de motivación. Así, Couch y Keniston (1960, 1961) creían que los individuos que respondían a una pregunta, llegaban a una respuesta que reevaluaban para evitar la desaprobación o una imagen negativa de sí mismos. Esto provocaría una latencia de respuesta superior a las de aquellas respuestas que no se reevaluasen. Otros autores, sin embargo (Scklenker 1980, Leary y Kowalski, 1990; Schneider, 1981) tenían una posición contraria. Según ellos, las respuestas aquiescentes serían mucho más rápidas que las respuestas contestadas apropiadamente, ya que no habría una búsqueda introspectiva del sujeto de información relevante.

Frente a las teorías motivacionales, existen las teorías cognitivas, que consideran este tipo de respuesta como un problema de los procesos cognitivos del respondiente. Autores como Zucherman et al., (1995) plantean la posibilidad de que el problema se debe a una búsqueda sesgada de la información debida, a su vez, a un problema de flexibilidad cognitiva. Es decir, ante una pregunta, un individuo simplemente piensa en uno o dos ejemplos confirmatorios antes de responder en lugar de hacer todo un proceso cognitivo en busca de evidencia que lo contradiga.

Un modelo, que podría entenderse como mixto, viene de la mano de Gilbert (1991), el cual describió un proceso de creencia "Espinoziano" en dos etapas: una de comprensión y otra de reconsideración. En la primera etapa se aceptaría la premisa sin ningún tipo de esfuerzo, mientras que en la segunda habría una reconsideración de la declaración a la luz de la evidencia donde se toma la decisión de rechazar o aceptarla. Esta última etapa depende tanto de la capacidad como de la motivación para hacer el esfuerzo. Un sujeto aquiescente seria pues aquel que tras la primera etapa no reconsidera ni reevalúa la afirmación que se le hace. Knowles y Condon (1999) encontraron evidencia adicional consistente con la propuesta de Gilbert, observando que las personas que respondían de forma aquiescente a los ítems eran más rápidos que las que contestaban apropiadamente.

1.2.Características personales vs itemétricas

Ahora bien, conociendo las teorías que describen el proceso por el cual una persona responde de forma aquiescente, los investigadores también se interesaron en ver qué características personales e itemétricas hacían que los test fueran contestados de forma aquiescente.

Características de personalidad

La ACQ se ha relacionado con rasgos de personalidad como la inteligencia (educación y cultura), la impulsividad y la extraversión. En concreto, la aquiescencia parece correlacionar negativamente con la inteligencia, de forma que las personas menos inteligentes o con un nivel de educación menor suelen ser más aquiescentes (Forehand, 1962; Gudjonsson, 1990; Meisenberg y William, 2008; Navarro et al. 2018; Soto, et al., 2008, 2011). Así mismo, existe una relación positiva con la extraversión y la impulsividad (Couch y Keniston, 1960). Observándose también que la respuesta aquiescente es más habitual en aquellas personas preocupadas e inseguras por los prejuicios que pudiera causar su respuesta (Ross y Mirowsky, 1984).

Diferencias culturales y sociodemográficas.

Bachman y O'Malley (1984) observaron que africanos e hispanos presentaban mayor ACQ que los americanos caucásicos. También se encontró una mayor tendencia de ACQ en sociedades menos individualistas (e.g. Japón, Latinoamérica) (ven Herk, et al., 2004).

Entendida como constructo, la aquiescencia presenta estabilidad y consistencia a medio plazo (Billliet y Davidov, 2008; Weijters, et al., 2010a; Weijters, et al., 2010b; and Wetzel et al, 2015). No obstante, la edad es un factor que hace que esta tendencia varíe. Se observa que los niños, adolescentes y ancianos tienen más tendencia a contestar de forma aquiescente que los adultos (Vigil, et al., 2013, 2015).

A pesar de haberse llegado a resultados consistentes en este apartado, debe tenerse en cuenta que las variables sociodemográficas son el tipo de diferencia individual que menos explicaría la presencia de sesgo (no más de un 5% según Weijters, et al., 2010a).

Características y situacionales e itemétricas

Una línea de investigación alternativa a las resumidas hasta ahora se ha centrado en determinar si el diseño y características del test o el tipo de presentación afectaría a los sujetos a la hora de responder a los ítems de forma sesgada. En cuanto a la presentación, un estudio realizado por Weijters et al., (2010b) observó que las entrevistas telefónicas suelen producir valores más altos de ACQ frente a las opciones online y de papel y lápiz.

Las características itemétricas se refieren a esas propiedades individuales de los ítems, como la discriminación o el formato del ítem. Angleitner et al. (1986) estudiaron las propiedades itemétricas de los cuestionarios de personalidad: comprensión de los ítems, ambigüedad, abstracción, la auto-referencia, o la evaluación que hacen los individuos sobre los ítems. Observaron que los ítems ambiguos (ítems que contenían negaciones, conjunciones o con más de una cláusula) tendían a elicitar más respuestas aquiescentes. Dada la definición que aportaron los autores a las propiedades estudiadas, la comprensión de los ítems también podría ser un factor de riesgo para la aquiescencia. Condon et al., (2006) se centraron en la complejidad y el tamaño del ítem confirmando de forma empírica las hipótesis previas que sugerían que aquellos ítems más largos y complejos (y, por lo tanto, con una alta probabilidad de ser menos comprensibles) tenían mayor probabilidad de desencadenar una respuesta aquiescente.

El número de opciones de respuesta no parece afectar en demasía al estilo de respuesta en el que nos centramos (Weijter, et al., 2010b), no obstante, cuando estas opciones de respuesta están completamente etiquetadas parece que la tendencia es mayor. Por otro lado, el formato de presentación del test también parece afectar al problema de la aquiescencia

1.3.<u>Efectos de la Aquiescencia</u>

Sabiendo que alguna de las causas por las que se produce la aquiescencia no son manipulables experimentalmente (v.g. rasgos de personalidad, cultura, socialización…), y que por lo tanto escapan a nuestro control, el planteamiento de buscar formas de detectar y controlar la aquiescencia debe ponerse sobre la mesa. Pero, ¿realmente es necesario ese control? ¿Qué efectos produce la aquiescencia cuando ésta no se controla?

Desde el marco teórico del análisis factorial, en el que se basa el presente trabajo, La aquiescencia no controlada afecta tanto a la estructura factorial como a las puntuaciones estimadas que se derivan de dicha estructura. En el nivel más bajo del análisis de ítems, distorsiona la correlación entre ítems sobreestimando las correlaciones cuando los ítems están redactados en la misma dirección, e infraestimándola cuando lo están en sentido opuesto (Paulhus y Vazire, 2007; Morales-Vives et al, 2017; Vigil-Colet, et al., 2020). Como consecuencia el análisis factorial (y otras técnicas multivariadas de reducción de datos) pueden dar lugar a dimensiones espurias que reflejen efectos de método en lugar de diferencia individuales en los rasgos, e incluso una mezcla de ambos (Winkler, et al., 1982). En el marco específico de una solución factorial, la consecuencia habitual de los efectos de método arriba descritos en un test que pretende medir una sola dimensión con ítems positivos e ítems revertidos, es la división artefactual en dos factores, compuestos cada uno por los ítems de cada dirección (Horan et al,, 2003; DiStefano y Motl, 2009). En suma, tenemos un problema de multidimensionalidad espuria, en la que hacen falta dos factores para obtener un buen ajuste a los datos, aun cuando, a efectos de contenido, el test sea unidimensional. Nótese que, (a) si solo se extrae un factor, el ajuste será deficiente, pero, (b) si se extraen dos, se alcanzará probablemente un ajuste aceptable, pero estos factores no tendrán interpretación en términos de contenido. Finalmente, cabe notar que estos efectos se producirán tanto en soluciones exploratorias como confirmatorias. (Morales-Vives, et al.,2017).

Además, de los efectos generales descritos arriba, deben tenerse en cuenta los posibles efectos de interacción con los rasgos de personalidad que se evalúan, ya que algunos de estos rasgos pueden estar relacionados con la tendencia a dar respuestas aquiescentes. Finalmente, debe tenerse en cuenta el posible efecto de replicabilidad a través de muestras si algunas de estas muestras presentan una tendencia acusada a decir que sí independientemente del contenido.

1.4.Control de la Aquiescencia.

Método de Balanceo

Conociendo los efectos que puede producir la presencia de Aquiescencia, cabe concluir que es necesario tener métodos de control de este estilo de respuesta. Sin embargo, la revisión de la literatura indica que el desarrollo de tales métodos no ha sido un camino fácil.

Uno de los grandes problemas iniciales en el desarrollo de los métodos de control de la respuesta aquiescente fue el de estimar qué proporción en la varianza de las respuestas reflejaba los efectos del rasgo y que proporción era debida al sesgo, ya que en escalas compuestas únicamente por ítems positivos esa información no se podía desentrañar tan fácilmente. En cambio, si se pudiera modificar una parte de los ítems de forma que estuviesen midiendo en la dirección opuesta del rasgo, se consideró que sería posible obtener puntuaciones "balanceadas" que estarían relativamente libres del impacto de la ACQ (Rundquist, 1966). Este es el método más clásico para el control de la aquiescencia (Baumgartner y Steenkamp, 2001; Nunnally, 1978; Vigil-Colet, et al., 2020) y a día de hoy se ha mantenido como una de las formas más seguras y satisfactorias de controlar el sesgo.

El concepto de ítems negativos, revertidos o invertidos puede ser confuso ya que puede entenderse desde puntos de vista confrontados. Esta confusión viene por el modo en el que el ítem pasa a medir el rasgo desde la dirección contraria. Por un lado, estarían los ítems que contienen partículas negativas en su enunciado y por otro los que se redactan siempre en positivo, pero se orientan hacia uno u otro polo del rasgo. Por ejemplo: en una escala de extraversión, los ítems: "Me gusta estar en compañía" y

"Me gusta quedarme solo en casa" están redactados los dos en positivo, pero miden direcciones opuestas del rasgo.

Los ítems con contenido negativo o redactados de forma negativa no siempre parecían ser la mejor solución. Taylor y Bowers (1972), observaron que un ítem formulado de manera negativa produce una respuesta media más alta que su contraparte formulada de manera positiva. En cuanto al diseño de instrumentos de medición, la inclusión de este tipo de ítems, puede resultar en respuestas menos precisas y, por lo tanto, afectar a la consistencia interna debido a la correlación ítem total baja y a la validez de los resultados obtenidos (Schriesheim y Hill, 1981; Ebesutani et al., 2012, Paulhus y Vazire, 2007). Wong, et al. (2003) informaron que los ítems formulados de manera negativa pueden comprometer la dimensionalidad de una escala, derivando al ya mencionado efecto de método, es decir, una estructura que teóricamente es unidimensional resulte en bidimensional ubicándose en factores diferentes los ítems positivos y negativos (Paulhus y Vazire, 2007; Weijters y Baumgartner, 2012).

Dados los problemas que provoca el uso de este tipo de ítems, actualmente se recomienda llevar a cabo el balanceo utilizando ítems redactados positivamente que miden la versión opuesta del rasgo y que podríamos llamar "antónimos". La ventaja de usar esta estrategia de redactar los ítems usando antónimos es que no requiere negaciones y eso probablemente facilite la comprensión de los sujetos a la hora de responder los ítems. Sin embargo, a pesar de que el procedimiento de balancear los ítems es una estrategia más que sensata, se debe tener cuidado con que los ítems antónimos se orienten de forma estrictamente opuesta a los ítems directos originales. Además, el uso de este método está limitado a rasgos que se pueden entender como bipolares. ¿Cómo podríamos redactar ítems sobre rasgos que se conceptualizarían mejor como unipolares?

Desde hace tiempo, el diseño y construcción de muchas escalas psicométricas ya implementan el método del balanceo, preocupándose por redactar ítems claros, evitando ítems redundantes y ambiguos (Krosnick, 1999).

Construcción de escalas específicas de Aquiescencia.

Un método alternativo al balanceo para controlar la aquiescencia, que tuvo cierta resonancia en las etapas iniciales, era construir una escala de aquiescencia independiente al test de contenido bajo la suposición de que la aquiescencia se debía principalmente al individuo y no a las características del test.

En una línea similar, Watson (1992) propuso la administración de una prueba adicional con ítems de palabras positivas y revertidas para calcular un índice que evaluaba el grado en que un encuestado tiende a ser aquiescente. Posteriormente, utilizando el análisis factorial confirmatorio, este autor propone un modelo con ayuda del índice de forma que la aquiescencia esté separada en un factor independiente. Las escalas de control basadas en los principios descritos hasta ahora se empezaron a desarrollar en la década de los sesenta, siendo las más conocidas las de Couch y Keniston (1960) y la de Hanley (1961). También se propusieron enunciados interrogativos que, según Wong et al. (2003) eliminaban los efectos derivados de la redacción del ítem. Algunos manuales sobre diseño de cuestionarios sugieren escalas de respuesta de elección forzada" o "constructo específico" (Kroskick, 1999; Krosnick et al., 2005, Krosnick y Presser, 2010; Pasek y Krosnick, 2010). Los cuestionarios de constructo específico son aquellos en los que se pregunta directamente (en vez de afirmar) sobre la dimensión subyacente. Por ejemplo, un ítem estándar sería "Me agrada hacer listas", mientras que una pregunta de constructo específico sería: "En qué medida te agrada hacer listas".

Otra línea de desarrollo fueron las pruebas "libres contenido". Se basaban en el principio de que los sesgos de respuesta podrían estudiarse al minimizar el contenido verbal mediante la inclusión de ítems sin sentido. Es decir, si el enunciado de un ítem sin sentido no afecta directamente a una respuesta particular, entonces la respuesta dada a ese ítem puede deberse a otros factores como la aquiescencia (Cronbach, 1946). Este sentido, Cruse (1966) usó ítems escritos en árabe sin sentido preguntando a los participantes el grado de acuerdo o desacuerdo, con la idea de que las respuestas que darían no serían aleatorias, de forma que si se observase cierta tendencia sistemática de respuesta, ésta sólo puede deberse a la aquiescencia.

Método de Ipsatización

El método de la ipsatización se propuso como una alternativa a los problemas que causaban los ítems redactados negativamente y que se han discutido antes (Wong et, al., 2003). Aunque existen variantes, el principio general es el de eliminar la media (Clemans, 1968) dejando así sólo aquella varianza que se puede atribuir a la dimensión. Por ejemplo, se pueden calcular restando la puntuación media del test a todas las puntuaciones de ese individuo (Cattell, 1944). Ten Berge (1999) examinó las propiedades de la transformación ipsativa en escalas balanceadas. Al calcular las puntuaciones de los sujetos se debían invertir aquellas puntuaciones de los ítems revertidos y luego calcular las puntuaciones promedio de los ítems. Por lo tanto, debido a la inversión, las puntuaciones no serían completamente ipsativas ya que quedará varianza entre sujetos. En palabras de Primi, Santos et al. (2019), la corrección ipsativa es una acción de desglose que elimina la varianza de la aquiescencia de las puntuaciones de los ítems.

En la práctica, sin embargo, la ipsatización puede tener algunas desventajas. En particular, al aplicar estos ajustes, se pierde la información relativa a las posibles diferencias inter-individuales (Dolnicar y Grün, 2007; Fischer, 2004, Hicks, 1970). Esto puede ser problemático en ciertos análisis estadísticos y puede conducir a la eliminación de diferencias válidas entre los participantes. Por otro lado, la validez en el nivel de los individuos puede ser dudosa; ya que, aunque se podría esperar un aumento de la validez al usar el procedimiento, la corrección acarrea una reducción de las puntuaciones extremas acercándolas al punto medio

En 1993, Chan y Bentler (1993) desarrollaron un procedimiento para ajustar modelos de factores a datos que ya poseen una estructura ipsativa. Aplicando el método a los datos ya ipsatizados, se eliminaría aún más el sesgo de la estructura de covarianza. El método requiere que existan varios factores para que el modelo esté identificado, no obstante, Savalei y Falk (2014) realizaron una adaptación de forma que para ajustar el modelo eliminaban una variable (que luego obtenían a partir de la suposición de que la suma de las cargas era igual a 0) a los datos ya ipsatizados, ajustando el modelo de 1 factor con el resto de variables.

Métodos Post-Hoc

Frente a los métodos anteriores, quizá más ligados a una interpretación de la aquiescencia como rasgo de personalidad y no como sesgo per sé, se encuentran los métodos de control estadístico post hoc.

No obstante, antes de entrar en las correcciones post-hoc basadas en la conceptualización de sesgo, es interesante comentar brevemente el estudio de Johanson y Osborn, (2004) donde estas correcciones se aplican a las personas. Los autores sugieren que la ACQ se definiría operativamente como funcionamiento diferencial de la persona (FDP) entendida como la observación de una respuesta diferencial de una persona a dos grupos de ítems (por ejemplo, ítems positivos y negativos). La conclusión a la que llegan es que: igual que se eliminan aquellos ítems con funcionamiento diferencial del ítem, en la construcción de escalas deberían no tenerse en cuenta a las personas que respondan de forma aquiescente.

En 1958, Webster sugirió un método de corrección basado en la regresión lineal. Primero utilizaba índices de frecuencia para cuantificar el grado en el que un individuo empleaba un estilo de respuesta. El método, a grandes rasgos, consistía en calcular el residuo de regresión restando la puntuación esperada de la observada. El resultado era la eliminación o minimización de la influencia del sesgo en las inter-correlaciones entre ítems y la mejora de la precisión de los análisis de esas correlaciones. El procedimiento permitía obtener una matriz de correlación parcial de primer orden en la que el impacto de la ACQ se había parcializado de las correlaciones de primer orden. Dicha matriz "limpia" servía después como entrada a los análisis multivariados posteriores (Winkler et al., 1982). A pesar de poder implementarse fácilmente, asumía que la relación entre el rasgo y el estilo de respuesta es lineal, hecho que no siempre es cierto (Wetzel et al., 2016).

Mirowsky y Ross (1991) fueron los primeros en modelar la aquiescencia por medio de la introducción de un factor con cargas fijas o restringidas que afectaría a todos los ítems de una escala (véase figura 1). A partir de este, se han propuesto otros modelos similares como el del Billiet y McClendon, 2000; Cambré et al., 2002; Watson, 1992; Welkenhuysen-Gybels et al., 2003 o el de Maydeu-Olivares y Coffman (2006).

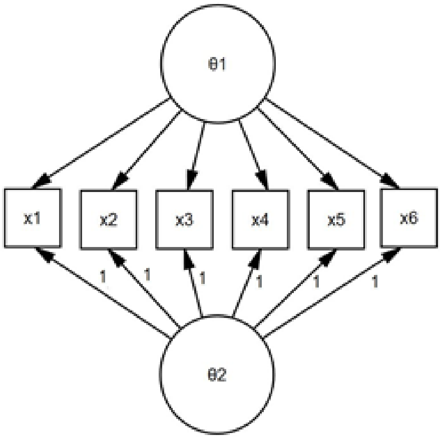


Figura 1. Modelo bidimensional con un factor de contenido (θ_1) y otro de aquiescencia (θ_2). las cargas del factor de contenido son libres, mientras que las del factor de ACQ son restrictas.

El modelo de intercepto aleatorio propuesto por Maydeu-Olivares y Coffman (2006) plantea modelar las diferencias individuales sistemáticas que no se pueden capturar por el factor de contenido relajando la suposición de que el intercepto debe ser común para todos los encuestados. La ventaja de este enfoque es que no hay restricción en cuanto al balanceo de los ítems o una posible diferencia en el valor de las cargas factoriales entre los ítems positivos y negativos (Savalei y Falk, 2014). No obstante, asume tau-equivalencia en las cargas de los ítems en el factor de aquiescencia (de nuevo pues, tenemos restricciones en este factor). Este es un supuesto difícil de cumplir, ya que algunos ítems pueden ser más propensos a generar respuestas de aquiescencia que otros debido a su formulación, contenido o contexto.

Frente a los modelos confirmatorios, mucho más restrictivos, se han desarrollado varios métodos exploratorios. Desde esta perspectiva, Ferrando et al. (2003) propusieron un método para ajustar el modelo de aquiescencia como factor latente. El método tiene su origen en el trabajo de ten Berge (1999) el cual, propuso un procedimiento basado en componentes principales para eliminar la varianza debida a la aquiescencia. El método exploratorio de factores no restringidos (Ferrando et al., 2003) se diferencia de su predecesor al cambiar el enfoque basado en componentes principales por un análisis de factores no restringidos. La desventaja más evidente de este modelo es la falta de robustez cuando se viola el supuesto de balanceo en las cargas de contenido. No obstante, aun siendo diseñado para escalas completamente balanceadas el método es bastante eficaz para estimar las cargas de contenido y aquiescencia incluso cuando la escala está tan solo parcialmente balanceada en contenido, y las cargas en el factor de aquiescencia son heterogéneas, lo que supone una ventaja sobre los modelos restringidos.

A fin de eliminar una de las desventajas del modelo anterior Lorenzo-Seva y Ferrando (2009) elaboran un método, basado en el procedimiento de Watson (1992) que elimina la varianza debida a la aquiescencia en escalas parcialmente balanceadas. Utilizando el análisis factorial exploratorio como base para el método, el primer paso y característico del método es seleccionar un núcleo centroide de ítems balanceados lo más grande posible que contendría todos los ítems en dirección inversa y el mismo número de ítems en dirección directa, eligiendo aquellos más afectados por la aquiescencia. Este procedimiento no estaría afectado por las limitaciones de sus "grandes competidores" pudiéndose obtener una matriz de cargas de la aquiescencia mucho más ajustada y siendo un método robusto ante escalas no balanceadas. No obstante, no sería recomendable si el subconjunto centroide es demasiado pequeño, dicho de otra forma, si hay un desequilibrio muy alto, es probable que no se pueda definir el factor de aquiescencia.

Autores como Aichholzer (2014), por su parte, se han decantado por desarrollar modelos híbridos. El modelo RI-EFA estima libremente las cargas factoriales, procedimiento típico de los modelos exploratorios y lo combina con una segunda etapa restringida típica de los modelos de CFA en que las cargas factoriales del ítem en el factor de aquiescencia están restringidas siendo tau-equivalentes, lo que supone una limitación similar a los procedimientos confirmatorios. Por otro lado, Aichholzer no deja claro hasta qué punto el método es sensible al desequilibrio de los ítems.

1.5.El problema de la validez

Actualmente, la validez es considerada un concepto unitario y multifacético. Es decir, para hablar de la validez de una prueba se deben unir múltiples tipos de evidencia, las cuales no son alternativas sino complementarias (Messick, 1990). Hay diferencias entre los autores a la hora de clasificar los tipos de validez, pero en este trabajo seguiremos la propuesta de Messick (1990) en tres grandes grupos: validez de contenido, validez relacionada con el criterio y validez de constructo. La primera, trata de saber si el contenido de la prueba muestra la información sobre la que se van a hacer inferencias posteriormente. La validez de criterio analiza las relaciones de las puntuaciones de la prueba con variables externas llamadas criterios. Y la de constructo se fija en las cualidades que mide una prueba.

Siendo consciente de que para realizar un estudio completo de validez se requieren incluir diversos tipos de pruebas, en el presente trabajo y a fin de acotar el campo de estudio, se profundizará aquí en la validez externa o relacionada con el criterio. Además, diversos autores consideran que, en concreto, obtener una buena relación de este tipo se considera un indicativo de la calidad de la medición (Jenkinson et al., 1994; Malhotra y Krosnick, 2007).

La validez de criterio es entendida como la capacidad que tienen las puntuaciones en un test para relacionarse con variables externas al mismo que, teóricamente, tienen relación con el constructo que mide el test. Estas variables pueden ser las medidas en un criterio objetivo o las puntuaciones en otros tests que miden los constructos teóricamente relacionados. Existen muchas posibles medidas de criterio y diversos contextos en los que se puede usar la prueba, por lo tanto, la validez externa depende del criterio que se use, y técnicamente habría tantas evidencias de validez externa como "criterios". La ventaja es que la validez externa permite tanto examinar las relaciones con otras variables que no forman parte directa de la prueba en sí (tal y como la definición indica), el grado de generalización de la prueba e incluso las limitaciones o fuentes de invalidez. Por otra parte, la idea de que la validez relacionada con el criterio depende del propio criterio con el que se lo compara es un tema controvertido, ya que la mayoría de las veces este criterio es otra prueba y ésta puede ser una medida imperfecta o sesgada. Por lo tanto, a la hora de utilizar un criterio se debe comprobar que ésta variable sea válida, relevante y confiable ya que es fundamental para asegurar que las conclusiones sean sólidas.

En lo que respecta a la ACQ, la gran mayoría de la literatura existente se ha concentrado en su impacto a nivel de características estructurales e internas (validez de constructo en la clasificación de

Messick). En cambio, los estudios que extienden este impacto a las relaciones con variables externas relevantes son mucho más escasos (Primi, De Fruit, et al., 2019). Además, estos pocos estudios se centran en un solo tipo o evidencia y aparecen de manera dispersa, poco clara y sin una línea de investigación aparente que los integre de manera coherente. En el presente trabajo nos centraremos en dos estudios que se dedican a observar el grado de validez en el caso del desbalanceo de los ítems y el efecto en la validez de distintos métodos de corrección de la aquiescencia.

Soto y John (2019) llevaron a cabo una investigación centrada en analizar cómo ciertas características internas de las escalas de medición de personalidad, específicamente la longitud, la amplitud y el balanceo de los ítems, impactan en la validez externa. Brevemente, se puede decir que, en relación a la longitud de la escala, se observó que, en un primer momento, existe una relación positiva entre el número de ítems en una escala y la precisión de la medición. Sin embargo, se identificó un punto óptimo en el que los beneficios en términos de precisión comienzan a disminuir. En lo que respecta a la amplitud de la escala, se encontró que las escalas de personalidad que abarcan una gama más amplia de contenido relacionado con el rasgo tienden a tener asociaciones de validez externa más fuertes. Esto se debe a que estas escalas son capaces de capturar una mayor cantidad de información no redundante sobre la personalidad.

En lo que respecta al balanceo de los ítems, se destacó que la tendencia de algunos individuos a responder de manera aquiescente puede sesgar las asociaciones de validez entre los rasgos de personalidad y los criterios externos. Este sesgo es especialmente relevante cuando tanto las escalas de rasgos como los criterios presentan un desbalanceo en la cantidad de ítems positivos y negativos ya que puede llevar a sesgos de validez positivos o negativos. Además, cuando mayor sea este grado de desbalanceo más fuerte será el sesgo.

Los sesgos de validez positivos se producen cuando ambos, la medida de contenido y la variable externa tienen un desbalanceo en la misma dirección (por ejemplo: ambos tienen más ítems positivos que negativos). En este caso, las correlaciones son más altas de lo que deberían ser en ausencia de sesgo. Por su parte, el sesgo de validez negativo implica que las correlaciones sean más bajas de lo que deberían ser en ausencia de aquiescencia y se produce cuando el desbalanceo entre las dos medidas se produce en direcciones opuestas (una de las escalas tiene más ítems positivos y la otra tiene más ítems negativos).

El segundo estudio a revisar es el de Scharl y Gnambs (2022) en la cual observaron el impacto que tenían los métodos de corrección en la validez externa. Para elegir los criterios más relevantes, se basaron en investigaciones previas, la literatura y la teoría subyacente a los constructos de autoestima y necesidad de cognición. A partir de ahí, compararon los modelos corregidos (por sesgo) y no corregidos. Para corregir la respuesta aquiescente se usaron tres métodos: (a) indicadores de estilos de respuesta que se derivaban de ítems independientes a los que componían el test, (b) el modelo multidimensional de crédito parcial generalizado (MGPCM), que permite modelar los sesgos como características de las personas y (c) la técnica de árboles de procesamiento multinomial, que se utiliza para modelar el proceso de toma de decisiones y permite modelar varios sesgos de aquiescencia como etapas de decisiones en el proceso de respuesta. Los resultados observados encontraron que los métodos de corrección tenían un impacto en las correlaciones entre el contenido evaluado por el test y los criterios. Y que, además, las correlaciones variaban en función del tipo de método de control de aquiescencia que se utilizaba: el tamaño e incluso el signo de la correlación entre el rasgo de contenido y una variable externa variaban en función del método. Estos resultados abren pues una puerta para poder seguir investigando el efecto del control de la aquiescencia en la validez externa, en la línea que planteamos a continuación.

Supongamos que tenemos un test A libre de sesgos que deseamos relacionar con una variable externa relevante, digamos las puntuaciones en un segundo test B (no es esta la mejor estrategia de validación, pero sí la más frecuente). Desde el punto de vista del modelo factorial, un método muy simple para estimar el coeficiente de validez es calcular la correlación entre las puntuaciones factoriales estimadas del test A y las del test B. Esta correlación sería un coeficiente de validez empírico El correspondiente coeficiente de validez teórico se definiría como la correlación entre las puntuaciones factoriales "verdaderas" en ambos tests (véase Lord y Novick 1968). Si, por simplicidad, nos limitamos a la inferencia psicométrica (prescindiendo de la distinción muestra-población de individuos), se deduce, tanto del modelo factorial como de la teoría del test en general, que el coeficiente de validez empírico es un estimador atenuado del coeficiente teórico. La atenuación se debería a la presencia inevitable de la varianza de error tanto en un test como en otro. Adviértase aquí que, si la validación se llevara a cabo contra un criterio objetivo, la única fuente de error serían las puntuaciones en el test.

Ahora, además, pensemos en la posibilidad de que el test A tenga un sesgo de ACQ (u otro estilo de respuesta). En ese caso, la magnitud de la atenuación variará ya que las puntuaciones factoriales de nuestro test A presentan, además de la varianza debida al rasgo, la varianza debida a la aquiescencia. Y, además, ambas fuentes de varianza son sistemáticas en contraste con la de error que es aleatoria. Los resultados descritos anteriormente y el funcionamiento del propio modelo, permiten pensar que la atenuación será menor o mayor dependiendo del tipo de sesgo, la dirección esperada en la relación entre el rasgo y la variable externa y la dirección del desbalanceo de las escalas (misma dirección, sesgo positivo; dirección opuesta, sesgo negativo)

2. Dependencia Local

Consideremos una escala unidimensional formada por una pareja de ítems. Al ser unidimensional ambos ítems deberán medir tan solo la dimensión común que intentan medir. Por lo tanto, y si están bien diseñados, cabe esperar que los ítems correlacionen debido a que sus puntuaciones están (parcialmente) determinadas por dicha dimensión común (de hecho, este es el principio de Spearman en que se basa el AF). Además, de acuerdo con el principio de independencia local, que es común tanto en AF como en la teoría del test, la correlación entre los ítems se debe únicamente a la influencia de esta dimensión común; el resto de determinantes se pueden considerar como error aleatorio de medida. Por tanto, si se eliminase la influencia del factor común, estos ítems deberían dejar de estar relacionados. En el caso del AF, el principio de eliminar la influencia común se basa en el modelo lineal, y, se podría enunciar de forma algo más técnica como sigue: La correlación parcial entre dos ítems que miden una dimensión común cuando se parcializa la influencia de dicha dimensión debe ser cero. En el caso de la teoría de respuesta al ítem, el principio es más fuerte y se refiere a la probabilidad condicional. Sin embargo, en este trabajo sólo vamos a considerar el principio débil o lineal de independencia local (DLd). Desde esta perspectiva, las violaciones de la DLd pueden ser consideradas como errores de especificación, es decir, que el modelo estadístico que está utilizando para analizar los datos no refleja la estructura de las relaciones subyacentes.

Consideremos ahora que parte de la correlación observada entre nuestros dos ítems se debe a efectos de método que, además, no tienen nada que ver con la dimensión a medir (por ejemplo, el contenido de los ítems es redundante, o los participantes tienden a aceptarlos ambos tengan el contenido que tengan). A este fenómeno se le conoce como varianza compartida de método, dependencia local (DL desde la TRI) o correlación residual (desde el modelo AF basado en el principio lineal o débil de independencia local).

2.1.Modelos y causas de dependencia local

Chen y Thissen (1997) diferencian entre dos modelos de dependencia local. El primero de ellos es el de dependencia local subyacente (Thissen et al., 1992) el cual se basa en la idea de una variable común adicional no modelada (generalmente menor) distinta del rasgo de contenido que se pretende medir. El otro modelo es el llamado dependencia local superficial (Thissen et al., 1992) y se manifiesta cuando los participantes responden de manera idéntica a dos ítems porque éstos son similares por diversas causas tales como formulación, similitud en el enunciado o ubicación en el instrumento.

Dentro del segundo modelo, la redundancia es, según Bandalos (2021), la razón que más frecuentemente se da en la literatura aplicada para explicar la aparición de residuos correlacionados (por ejemplo, Harry y Crea, 2018; Kim y Kamphaus, 2018; Kopp et al., 2011; Leong et al., 2018; Li et al., 2018; Napolitano y Job, 2018; Pincus et al., 2009; Snarr et al., 2009; Tovar et al., 2009; visto en Bandalos 2021). Pero no siempre es así. En primer lugar, los ítems redundantes no son, ni de lejos, la única causa de la aparición de los residuales correlacionados, sino símplemente una de ellas. Además, recurrir a la redundancia como justificación post-hoc para liberar residuales y mejorar el ajuste puede ser problemático (Cole et al, 2007).

En base a las posibles fuentes de correlaciones residuales sugeridas en la literatura tales como: efectos de método (Cole et al., 2007; Fornell, 1983; Saris y Aalberts, 2003), efectos de orden (Cronbach y Shalvelson, 2004; Green y Hershberger, 2000) o redundancia (Cole et al., 2007; Saris y Aalberts, 2003), hemos intentado elaborar una figura sumario que contenga, de forma detallada, las principales causas de la dependencia local débil referidas a test de rendimiento típico (personalidad, actitud, motivación…), ya que son estas las que se considerarán en el trabajo (Tabla 1).

Bandalos (2021) se interesa por la influencia que pueden tener en las respuestas de los individuos, la similitud y el orden en el que se presentan los ítems. De debe aclarar que Bandalos incluye en el concepto de similitud tanto la redundancia temática como la léxica. Basándose en estudios previos (Tourangeau et al., 1989), menciona el término de "efectos de contexto" basado en la idea de que las actitudes y opiniones no están almacenadas de manera fija en la memoria. sino que se generan ante preguntas específicas.

Tabla 1. Principales causas de la dependencia local débil

Interferencia externa	Interrupciones, material defectuoso, procedimientos de administración incorrectos.	
Redundancia temática	Ítems que no agregan información adicional, cuyo formato o contenido es similar entre ellos y cuya inclusión se considera innecesaria.	Disfruto saliendo con mis amigos Me encanta pasar tiempo fuera con mis amistades.
Redundancia percibida	Ítems que teóricamente, o que, a nivel de expertos, miden dos aspectos diferenciables de un rasgo, pero para el individuo son redundantes.	Noto que me distraigo fácilmente Siento que me cuesta prestar atención
Redundancia Léxica	Ítems que utilizan palabras similares, aunque la idea que expresen sea diferente.	Me da miedo montar en trasporte público Me dan miedo las muñecas de porcelana
Fatiga	Cansancio y falta de motivación. Frecuente en pruebas largas y exigentes.	
Práctica	Exposición repetida de la prueba donde los ítems se encuentran en la misma posición relativa.	
Redundancia Contextual	Varios ítems están vinculados al mismo contexto. También puede ser producida debido a que el sujeto tenga un nivel de conocimiento inusual sobre el contenido.	Tengo problemas para organizarme en el trabajo Prefiero evitar problemas en mi entorno laboral

Citando a Tourangeau et al. (1989) sobre el efecto del orden, Bandalos, sugiere que si dos ítems son similares (ya sea por redundancia temática, por redundancia léxica, por ambas o por cualquier otro motivo) cuanto mayor sea su proximidad física dentro de la escala, mayor probabilidad habrá de que den lugar a un residual correlacionado; dicho de otra forma, el efecto de orden moderaría el grado de similitud. En su estudio descubre otros hechos que son de interés: (1) observa que cuando se presentan varios tipos de redundancia entre los pares de ítems, el residual correlacionado será más fuerte; y (2) que la redundancia temática parece influir menos que la léxica.

Citando a Tourangeau et al. (1989) sobre el efecto del orden, Bandalos, sugiere que si dos ítems son similares (ya sea por redundancia temática, por redundancia léxica, por ambas o por cualquier otro motivo) cuanto mayor sea su proximidad física dentro de la escala, mayor probabilidad habrá de que den lugar a un residual correlacionado; dicho de otra forma, el efecto de orden moderaría el grado de similitud. En su estudio descubre otros hechos que son de interés: (1) observa que cuando se presentan varios tipos de redundancia entre los pares de ítems, el residual correlacionado será más fuerte; y (2) que la redundancia temática parece influir menos que la léxica.

2.2.Efectos de la Dependencia Local

Cole et al., (2007) indican que un modelado incorrecto o insuficiente de las correlaciones residuales puede llevar a sesgos en las estimaciones de los parámetros del modelo, errores en la interpretación de los resultados y posible distorsión en las relaciones entre variables. También puede haber problemas de ajuste del modelo, ya que no representará con precisión la estructura de los datos observados y por ende no podrá generalizarse a poblaciones más amplias o a diferentes contextos.

Los posibles efectos de la correlación residual no modelada sobre los índices de bondad de ajuste indican que el chi cuadrado podría no detectar la presencia de errores correlacionados (Kolenikov, 2011). Por otro lado, la raíz media cuadrática de los residuales estandarizados (SRMR), al ser un promedio, no sería sensible a los casos en los que la matriz residual incluya muchos ceros y valores muy pequeños y solo una pequeña proporción sean residuos grandes (Shi, et al., 2018).

A nivel de las puntuaciones individuales estimadas, DeMars (2021) desde la TRI encuentra que la presencia de dependencia local débil tiene un impacto en la interpretación de las puntuaciones estimadas. Esto es debido a que los parámetros sesgados pueden llevar a una interpretación incorrecta del rasgo que mida la escala.

Desde la TRI, los efectos de desdeñar la dependencia local son altamente conocidos. Lucke (2005) estudio el efecto de las correlaciones residuales con la consistencia interna desde el modelo de Rasch y observó que la dependencia local tiende a sobreestimar la fiabilidad.

Marais y Andrich (2008) llevaron a cabo un estudio en el que variaron la intensidad de la dependencia entre ciertos ítems seleccionados. Descubrieron que a medida que la intensidad de la dependencia aumentaba, la similitud en las respuestas también lo hacía. Lo que resultaba en que las estadísticas de ajuste y los parámetros mostraban discrepancias cada vez mayores en comparación con sus valores teóricos. Observaron que, con el aumento de la dependencia de las respuestas, la fiabilidad aumenta sistemáticamente debido al aumento en la similitud de las respuestas entre los ítems. Pero en el caso de la validez se aplica la paradoja de la atenuación de la teoría tradicional del test. Esta paradoja se refiere a la relación entre la fiabilidad y la validez de una prueba. Establece que a medida que aumenta la consistencia de la prueba, la validez aparente de la prueba tiende a disminuir ya que se vuelve menos sensible a las diferencias reales en el constructo.

2.3. *Métodos de detección de la dependencia local*

Dado que la dependencia local viola uno de los dos supuestos fundamentales de la TRI y esta violación puede tener efectos substanciales, el tema ha sido objeto de extensa investigación desde este marco teórico tanto con respecto a las consecuencias como a la propuesta y desarrollo de métodos para detectarla. Antes de centrarnos en aquellos más relevantes para este trabajo mencionaremos muy brevemente algunos de los más interesantes de tipo más general como el índice Q3 de Yen (1984) que estima correlaciones de primer orden entre los residuos de variables observadas; el estadístico G2 basado en modelos de TRI desarrollado por Chen y Thissen (1997) y la prueba de Mantel-Haenszel (MH) (Cochran, 1954; Mantel y Haenszel, 1959) adaptada por Ip (2001) para detectar desviaciones de la independencia local tanto entre pares de ítems y como entre grupos de ítems.

Los métodos de detección de residuales correlacionados desde los modelos factoriales son, quizá, menos sofisticados. La detección de residuos correlacionados en modelos de análisis factorial exploratorio se basa en la inspección de la matriz de covarianza residual. Esta técnica, pese a que es una de las más extendidas tiene el inconveniente de no contar con el más que plausible “desplazamiento” de la varianza residual que tendería a sobreestimar las cargas factoriales de los ítems involucrados en el doblete.

En un estudio de simulación llevado a cabo por la autora para el 9th European Congress of Methodology (2021), se buscaba conocer la probabilidad de ocurrencia de que la varianza residual no fuera visible en la propia matriz de covarianza residual. Dicho de otro modo, se buscaba ver en qué situaciones hacer un estudio de residuales basándose únicamente en el estudio de la matriz de correlaciones era insuficiente. Los resultados indicaron que el 48% de las veces el residual no se detecta a partir de la matriz de correlación: casi un 15% de las ocasiones, el residual "se propaga" hasta las cargas factoriales y un 33% la varianza residual se dispersa a lo largo de toda la matriz residual.

El uso de los índices de modificación (Sörbom, 1989) en los modelos de análisis factorial confirmatorio son la otra cara de la moneda de las herramientas para controlar los residuales en los análisis factoriales. A través de ellos se indican los posibles elementos no nulos en las celdas situadas fuera de la diagonal de la matriz de covarianza residual. Una vez detectados, se procedería a liberar los parámetros de covarianza residual y así lograr un mejor ajuste.

El método de la covarianza bayesiana de LASSO (least absolute skrinkage and selection operator) conocido como bayessian lasso confirmatory factor analysis (BLCFA) (Zhang et al., 2021) se desarrolló como alternativa al procedimiento derivado de los índices de modificación que funcionan de forma secuencial (es decir la detección y liberación de parámetros es uno a uno). En el BLCFA se estima simultáneamente toda la estructura de covarianza residual no nula bajo la única restricción de que la matriz sea positiva definida.

3. *Relación entre Aquiescencia y dependencia local débil*

Hasta ahora el lector ha podido apreciar que tanto la aquiescencia como la dependencia local son temas ampliamente estudiados y discutidos en la literatura. Ambos se pueden clasificar como determinantes al margen del contenido cuya presencia causa efectos relevantes, tanto a nivel estructural como a nivel de puntuaciones, en el proceso de medición. La aquiescencia es, probablemente, el estilo de respuesta prototípico mientras que la dependencia local se podría considerar más propiamente como un efecto de método. No es, además, descabellado suponer que ambos determinantes puedan actuar en forma conjunta en aplicaciones reales, probablemente con más frecuencia de lo que nos gustaría. Sin embargo, la literatura científica parece no haberse interesado hasta ahora en esta situación.

Imaginemos un modelo unidimensional en el cual, entre los ítems 1 y 2 existe un efecto de dependencia local débil que se manifiesta en una correlación residual. Desde el modelo factorial exploratorio, si forzamos a que la correlación residual se estime lo más cerca de cero posible, es muy probable que esta covariación no modelada se propague o reasigne dando lugar a sesgos en la estimación de los parámetros estructurales del modelo: cargas factoriales y, en el caso más general, correlaciones entre factores. Siguiendo con el caso más simple, el unidimensional, cabría esperar que el sesgo se manifestara como un sesgo de expansión, inflando las cargas de los dos ítems implicados, probablemente a costa de atenuar (o desinflar) las cargas de los no implicados. Ahora bien, ¿Qué se esperaría que ocurriese en un modelo bidimensional en el que se especificaran un factor de contenido y un factor de aquiescencia? En la figura 2 se muestra una modelización para este caso.

Si observamos la figura 2, si se produjese el efecto de propagación o reasignación, el residual podría "desplazarse" bien al factor de contenido bien al de aquiescencia. Por tanto, existe la posibilidad de que el factor de aquiescencia absorba parte de la correlación residual no modelada como tal. Esto plantea una serie de cuestiones que son la base de este trabajo:

¿Qué ocurrirá al controlar la aquiescencia cuando coexistan de forma simultánea la presencia de residuales correlacionados y la respuesta aquiescente? ¿es necesario un procedimiento mixto de control en el que se controlen conjuntamente la aquiescencia y dependencia local?

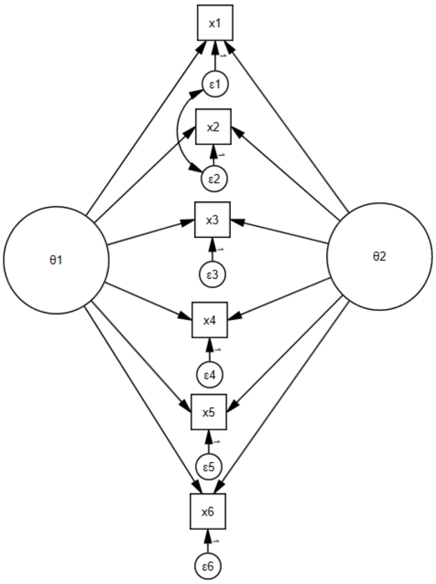


Figura 2. Modelo bidimensional con un factor de contenido (θ_1) y otro de aquiescencia (θ_2) y presencia de residual correlacionado entre los ítems x1 y x2.

Si se diera el caso de que la varianza residual fuera siempre absorbida por el factor de aquiescencia, y los métodos de control de aquiescencia fueran capaces de recoger toda esa varianza una modelización posterior de los residuales podría llegar a ser innecesaria.

Sin embargo, si se tiene en consideración que la varianza residual solo es visible en la matriz de cargas factoriales alrededor de un 15% de las veces, esto nos seguiría dejando con una necesidad de especificar aquellos residuos correlacionados (con base teórica) de algún modo. Por lo tanto, ¿hasta qué punto es necesario un protocolo que nos indique el modo de actuar? A día de hoy, no se puede saber el tipo de propagación (si la hay) que puede tener un residuo, por lo tanto, es posible que se requiera un protocolo o guía de actuación a la hora de ajustar modelos con sesgos y correlaciones residuales.

Capítulo 2. Objetivos e Hipótesis

La revisión bibliográfica presentada en este documento nos aporta indicios de la presencia de lagunas de información. Si bien el objetivo principal es determinar el posible impacto de dos determinantes no asociados al contenido cuando se presentan de forma conjunta, para llegar a responder a esto se deben cumplir unos objetivos previos.

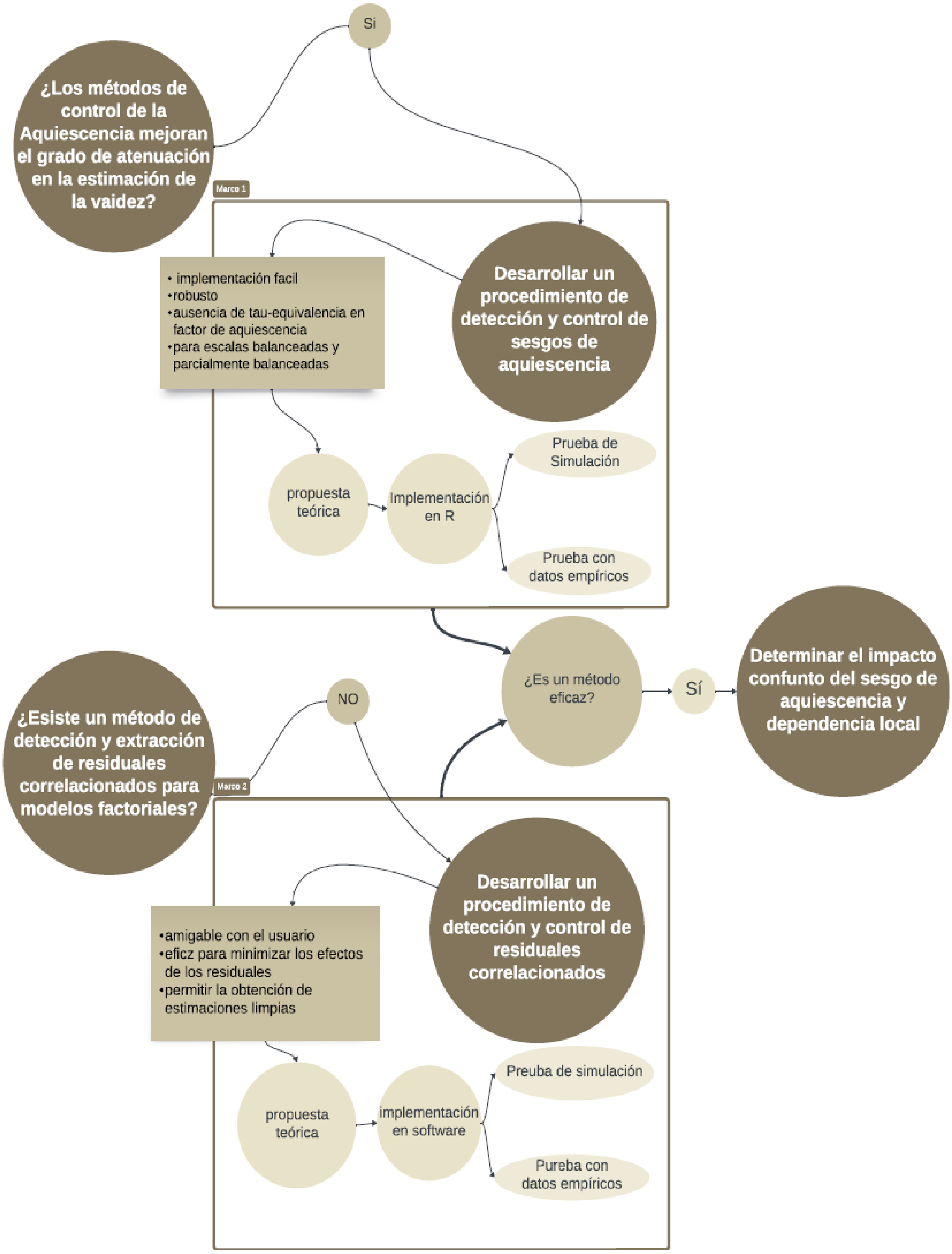
- **Objetivo 1:** Determinar el efecto que pueden tener los métodos de control de la aquiescencia en la validez externa de las escalas.
- **Objetivo 2:** Diseñar e implementar un nuevo método híbrido de control de aquiescencia que permita evaluar las propiedades estructurales del cuestionario y obtener estimaciones imparciales de la puntuación de los factores para cada encuestado.
- **Objetivo 3:** Diseñar e implementar un procedimiento nuevo (basado en propuestas previas) de detección y control de residuales correlacionados para modelos exploratorios que minimice los efectos de propagación de varianza cuando se fuerzan a cero los valores de la matriz de correlación
- **Objetivo 4:** Diseñar un procedimiento a través del cual se pueda evaluar el impacto conjunto del sesgo de aquiescencia y la dependencia local débil.

En consecuencia, una vez cumplidos los cuatro objetivos secundarios y el principal, obtendríamos:

- Evaluar el efecto de controlar la aquiescencia a través de modelos factoriales en la estimación de la validez. Un método de detección de aquiescencia, de fácil uso e implementado en R; cuya eficacia haya sido probada tanto en bases de datos simuladas como en datos empíricos.
- Un método de detección de dependencia local débil bajo el modelo de análisis factorial que permita minimizar los efectos de la covariación residual y obtener estimaciones más limpias de la parte estructural de la solución. Este método estará implementado en R y habrá sido probado con datos simulados y empíricos.
- Un procedimiento mixto a través del cual se pueda determinar el efecto de la aparición conjunta de dos tipos de determinantes no asociados al contenido.
- Un protocolo de actuación ante situaciones en las que se prevea una aparición conjunta de aquiescencia y dependencia local.

A continuación. se muestra un diagrama de flujo que mostraría gráficamente el procedimiento (véase figura 3).

Figura 3. Cronograma del procedimiento


Si
¿Los métodos de control de la Aquiescencia mejoran el grado de atenuación en la estimación de la vaidez?
Marco 1
Desarrollar un procedimiento de detección y control de sesgos de aquiescencia
• implementación facil
•robusto
•ausencia de tau-equivalencia en factor de aquiescencia
•para escalas balanceadas y parcialmente balanceadas
propuesta teórica
Implementación en R
Prueba de Simulación
Prueba con datos empíricos
¿Es un método eficaz?
SÍ
Determinar el impacto confunto del sesgo de aquiescencia y dependencia local
NO
¿Esiste un método de detección y extracción de residuales correlacionados para modelos factoriales?
Marco 2
Desarrollar un procedimiento de detección y control de residuales correlacionados
•amigable con el usuario
•eficz para minimizar los efectos de los residuales
•permitir la obtención de estimaciones limpias
propuesta teórica
implementación en software
Preuba de simulación
Pureba con datos empíricos

Capítulo 3. Métodos

Teniendo en cuenta que el proyecto está estructurado en 5 artículos, la metodología e instrumentos empleados son diversos. En la tabla 2 se muestra un resumen de los datos más relevantes de cada uno de los estudios. Además, en el capítulo 4, el lector podrá conocer la metodología de cada uno de los estudios en más profundidad.

No obstante, a continuación, se describen brevemente cada uno de los cuestionarios que se han empleado.

SAS: "Statistical Anxiety Scale (Vigil-Colet et al., 2008) Consiste en 24 afirmaciones que formarían un modelo jerárquico con un factor general que mediría ansiedad estadística y tres factores de segundo grado que miden: ansiedad por examen, ansiedad por demanda de ayuda y ansiedad por interpretación.

IDAQ: "Indirect-direct aggression questionnaire" (Ruiz-Pamies et al., 2014). Este test proporciona puntuaciones para los factores de agresión física, verbal e indirecta. Presenta control de deseabilidad social y de aquiescencia. Contiene 27 ítems en total: 12 directos, 11 antónimos y 4 son marcadores de deseabilidad social.

Raven: Matrices progresivas de Raven (Raven 1996). En su forma estándar consta de cinco conjuntos de doce matrices presentadas en blanco y negro. Aunque existen versiones abreviadas de nueve ítems (Bilker et al., 2012). Mide razonamiento abstracto y se considera una estimación no verbal de la inteligencia fluida.

WAIS: Wechsler Adult Intelligence Scale (Wechsler, 2003), prueba psicométrica que provee de cuatro puntuaciones: comprensión verbal, razonamiento perceptivo, memoria de trabajo y velocidad de procesamiento.

TPMAT: Thurstone's Primary Mental Abilities (Cordero et al, 1989). Se trata de una escala de inteligencia fluida y cristalizada que evalúa cinco aptitudes: comprensión verbal, espacial y numérica, razonamiento general y fluidez verbal.

PSYMAS: Psychological Maturity Assessment Scale (Morales-Vives, et al., 2013). Mide la madurez en la adolescencia y se compone de 27 ítems en escala Likert que se agrupan en tres factores: orientación del trabajo, autosuficiencia e identidad.

BAI: Belief in Astrology. (Chico y Lorenzo-Seva, 2006) Inventario en creencia en Astrología es un cuestionario unidimensional con formato Likert, formado por 24 ítems completamente balanceados

OPERAS: Overall Personality Assessment Scale (Vigil-Colet et al., 2013) contiene un total de 39 ítems y una estructura de 5 factores: Extraversión (7 ítems), estabilidad emocional (7 ítems), responsabilidad (7 ítems), amabilidad (7 ítems) y apertura a la experiencia (7 ítems). Además, permite controlar la aquiescencia y la deseabilidad social (4 ítems).

Tabla 2. Resumen de metodología procedimientos más relevantes de los artículos que se incluyen en esta Tesis.

		E1		E2	E3	E4	E5
Simulación	Diseño	Factorial		Factorial	Factorial	Factorial	Factorial
	Réplicas	100		200	200	500	500
	Software	R		R	Matlab	Matlab	R
	Análisis	ANOVA		ANOVA	ROC	ROC	ANOVA
Estudio Empírico	N	299	532	1309		743	2429
	Edad	18 y 60 (M =20.9; SD=4)	11 y 18 (M=14.75 y 2.1)	14 y 19 (M=16.4; SD=1.1)		18 y 60 (M=21; SD=4.3)	18 y 64 (M= 29.15, SD=14.65)
	Instrum	SAS	IDAQ, Raven WAIS TPMAT	PSYMAS		BAI	OPERAS
	Software	FACTOR	SPSS	R		FACTOR	R

Capítulo 4. Resultados

En este capítulo se muestran los 5 artículos que se incluyen en la presente Tesis. Estos artículos son:

Artículo Publicado: Hernández Dorado, A., Vigil Colet, A., Lorenzo Seva, U., & Ferrando, P. J. (2021). Is correcting for acquiescence increasing the external validity of personality test scores?. *Psicothema 33*(4), 639-646

Artículo en revisión: Navarro-Gonzalez, D.; Ferrando, P.J.; Morales-Vives, F. & Hernández-Dorado, A. SIREN: An Hybrid CFA-EFA R Package for Controlling Acquiescence in Restricted Factorial.

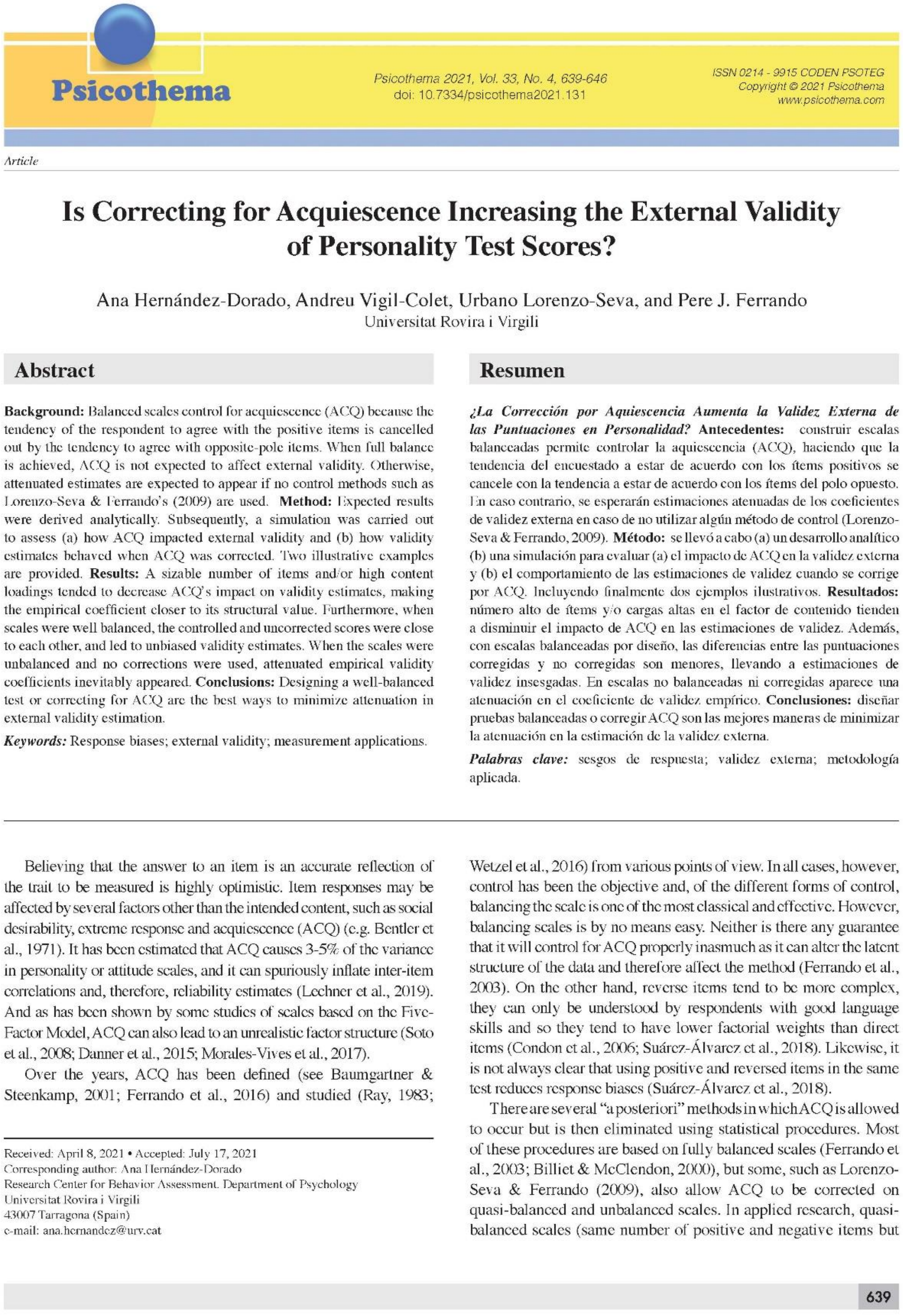
Paquete R: Navarro-Gonzalez, D., Ferrando, P. J., Morales-Vives, F., Hernandez-Dorado, A., & Navarro-Gonzalez, M. D. (2023). Package ‘siren’.

Artículo Publicado: Ferrando, P. J., Hernandez-Dorado, A., & Lorenzo-Seva, U. (2022). Detecting correlated residuals in exploratory factor analysis: New proposals and a comparison of procedures. *Structural Equation Modeling: A Multidisciplinary Journal, 29*(4), 630-638.

Artículo aceptado: Ferrando, P. J., Hernandez-Dorado, A., & Lorenzo-Seva, U. (in press). A Simple Two-Step Procedure for Fitting _Fully Unrestricted Exploratory Factor Analytic Solutions with Correlated

Residuals. *Structural Equation Modeling: A Multidisciplinary Journal*.

Artículo enviado: Hernández-Dorado, A. Ferrando, P.J. & Vigil-Colet. It’s not so bad! The impact and consequences of correcting for acquiescence when correlated residuals are present



Psicothema

Psicothema 2021, Vol. 33, No. 4, 639-646
doi: 10.7334/psicothema2021.131

ISSN 0214 - 9915 CODEN PSOTEG
Copyright © 2021 Psicothema
www.psicothema.com

Article

Is Correcting for Acquiescence Increasing the External Validity of Personality Test Scores?

Ana Hernández-Dorado, Andreu Vigil-Colet, Urbano Lorenzo-Seva, and Pere J. Ferrando
Universitat Rovira i Virgili

Abstract

Background: Balanced scales control for acquiescence (ACQ) because the tendency of the respondent to agree with the positive items is cancelled out by the tendency to agree with opposite-pole items. When full balance is achieved, ACQ is not expected to affect external validity. Otherwise, attenuated estimates are expected to appear if no control methods such as Lorenzo-Seva & Ferrando's (2009) are used. **Method:** Expected results were derived analytically. Subsequently, a simulation was carried out to assess (a) how ACQ impacted external validity and (b) how validity estimates behaved when ACQ was corrected. Two illustrative examples are provided. **Results:** A sizable number of items and/or high content loadings tended to decrease ACQ's impact on validity estimates, making the empirical coefficient closer to its structural value. Furthermore, when scales were well balanced, the controlled and uncorrected scores were close to each other, and led to unbiased validity estimates. When the scales were unbalanced and no corrections were used, attenuated empirical validity coefficients inevitably appeared. **Conclusions:** Designing a well-balanced test or correcting for ACQ are the best ways to minimize attenuation in external validity estimation.

Keywords: Response biases; external validity; measurement applications.

Resumen

¿La Corrección por Aquiescencia Aumenta la Validez Externa de las Puntuaciones en Personalidad? **Antecedentes:** construir escalas balanceadas permite controlar la aquiescencia (ACQ), haciendo que la tendencia del encuestado a estar de acuerdo con los ítems positivos se cancele con la tendencia a estar de acuerdo con los ítems del polo opuesto. En caso contrario, se esperarán estimaciones atenuadas de los coeficientes de validez externa en caso de no utilizar algún método de control (Lorenzo-Seva & Ferrando, 2009). **Método:** se llevó a cabo (a) un desarrollo analítico (b) una simulación para evaluar (a) el impacto de ACQ en la validez externa y (b) el comportamiento de las estimaciones de validez cuando se corrige por ACQ. Incluyendo finalmente dos ejemplos ilustrativos. **Resultados:** número alto de ítems y/o cargas altas en el factor de contenido tienden a disminuir el impacto de ACQ en las estimaciones de validez. Además, con escalas balanceadas por diseño, las diferencias entre las puntuaciones corregidas y no corregidas son menores, llevando a estimaciones de validez insesgadas. En escalas no balanceadas ni corregidas aparece una atenuación en el coeficiente de validez empírico. **Conclusiones:** diseñar pruebas balanceadas o corregir ACQ son las mejores maneras de minimizar la atenuación en la estimación de la validez externa.

Palabras clave: sesgos de respuesta; validez externa; metodología aplicada.

Believing that the answer to an item is an accurate reflection of the trait to be measured is highly optimistic. Item responses may be affected by several factors other than the intended content, such as social desirability, extreme response and acquiescence (ACQ) (e.g. Bentler et al., 1971). It has been estimated that ACQ causes 3-5% of the variance in personality or attitude scales, and it can spuriously inflate inter-item correlations and, therefore, reliability estimates (Lechner et al., 2019). And as has been shown by some studies of scales based on the Five-Factor Model, ACQ can also lead to an unrealistic factor structure (Soto et al., 2008; Danner et al., 2015; Morales-Vives et al., 2017).

Over the years, ACQ has been defined (see Baumgartner & Steenkamp, 2001; Ferrando et al., 2016) and studied (Ray, 1983; Wetzel et al., 2016) from various points of view. In all cases, however, control has been the objective and, of the different forms of control, balancing the scale is one of the most classical and effective. However, balancing scales is by no means easy. Neither is there any guarantee that it will control for ACQ properly inasmuch as it can alter the latent structure of the data and therefore affect the method (Ferrando et al., 2003). On the other hand, reverse items tend to be more complex, they can only be understood by respondents with good language skills and so they tend to have lower factorial weights than direct items (Condon et al., 2006; Suárez-Álvarez et al., 2018). Likewise, it is not always clear that using positive and reversed items in the same test reduces response biases (Suárez-Álvarez et al., 2018).

There are several "a posteriori" methods in which ACQ is allowed to occur but is then eliminated using statistical procedures. Most of these procedures are based on fully balanced scales (Ferrando et al., 2003; Billiet & McClendon, 2000), but some, such as Lorenzo-Seva & Ferrando (2009), also allow ACQ to be corrected on quasi-balanced and unbalanced scales. In applied research, quasi-balanced scales (same number of positive and negative items but

Received: April 8, 2021 • Accepted: July 17, 2021
Corresponding author: Ana Hernández-Dorado
Research Center for Behavior Assessment. Department of Psychology
Universitat Rovira i Virgili
43007 Tarragona (Spain)
e-mail: ana.hernandez@urv.cat

with unequal saturations) and partially balanced scales (different number of positive and negative items) are relatively common.

ACQ needs to be controlled and new forms of ACQ-control, such as Maydeu-Olivares & Coffman (2006), are still being investigated. The RIFA method, tested by De la Fuente & Abad (2020), could be a good alternative to the EFA-based method because it is easier to implement and is robust to the violation of the assumption of tau-equivalence in ACQ factor loadings. However, this approach is less accurate to highly heterogeneous ACQ loadings patterns because its estimate of the loadings of ACQ is the calculation of the average of these loadings. And, therefore, if we assume that the items will be affected differently by the ACQ and there is interest in the study of the ACQ factor, the Lorenzo-Seva & Ferrando (2009) model will be more appropriate. For more information on the development of ACQ control methods that have been developed, see Primi, Santos et al. (2019)

Although numerous studies have been made of ACQ, very few focus on validity, which might be partly because the concept is sometimes abstract and unfathomable. Primi, De Fruyt et al. (2019) observed differences in the criterion-related validity of the direct scores of the ACQ-uncorrected and corrected tests, and found "false-keyed" items to be more valid than "true-keyed" items.

Model-based Predicted Results

Basic Results and Validity Coefficients

We shall consider a general bi-dimensional model in which each item response is a measure of a content factor (θ_c) and an acquiescence factor (θ_a). The model has two parts: a measurement sub-model (1), and an extended structural sub-model (2) in which an external variable (a criterion) is regressed on the latent constructs defined in (1)

$$x_j = \lambda_{jc}\theta_c + \lambda_{ja}\theta_a + \varepsilon_j \tag{1}$$

$$y = \lambda_y\theta_c + \varepsilon_y \tag{2}$$

where λ_{jc} is the loading on content, λ_{ja} is the loading on ACQ and λ_y is the "true" validity coefficient. Both factors in (1) as well as the external variable (y) in (2) are assumed to be scaled which mean 0, and unit variance. In the simplest case: (a) ACQ and content are assumed not to be correlated, and (b) the criterion is an objective variable, and so uncorrelated with the ACQ factor.

This article will now go on to deal with the effects of ACQ on the coefficient of validity when the test scores are factor score estimates. Information about the corresponding effects when scores are raw or unit-weight sum scores will be provided by the authors on request.

By extending the definitions in Lord & Novick (1968, sect. 12.1), we now define the theoretical validity coefficient as the correlation $\rho\theta_c y$ which, given our adopted scaling and the fact that (2) is a single-regressor equation, is simply λ_y. Next, we define the empirical validity coefficient as the correlation between the content factor score estimates (based on model (1)) and y, denoted by "$\rho\hat{\theta}_{c'}y$". The relation between both coefficients is given by:

$$\rho\hat{\theta}_{c'}y = \frac{\rho\theta_c y}{\sqrt{1+\dfrac{1}{\Sigma_j \dfrac{\lambda_{jc}^2}{\sigma_{\varepsilon j}^2}}}} \tag{3}$$

where $\sigma_{\varepsilon j}^2$ is the error variance of the *jth* item. Equation 3 predicts that, if the factor estimates are corrected for ACQ, the empirical validity coefficient is still an attenuated measure of the "true" relationship between the test content and the criterion y (i.e. the theoretical validity). Clearly, attenuation is mitigated when test length and magnitude of the loadings increase, a result noted in previous studies (e.g. Soto & John, 2019).

Gone with the ACQ

Consider now that the measurement model (1) holds, and the data is fitted by the unidimensional model, so assuming that all the common variance is due to the content and that, therefore the presence of ACQ is ignored. Denote by $\hat{\theta}_g$ the maximum likelihood (ML) factor score estimates of the general factor. In this case, the validity relations are:

$$\rho_{\hat{\theta}_{g'}y} = \frac{\delta_c\rho\theta_c y}{\sqrt{1+\dfrac{1}{\Sigma_j \dfrac{\lambda_{jg}^2}{1-\lambda_{jg}^2}}}} \tag{4}$$

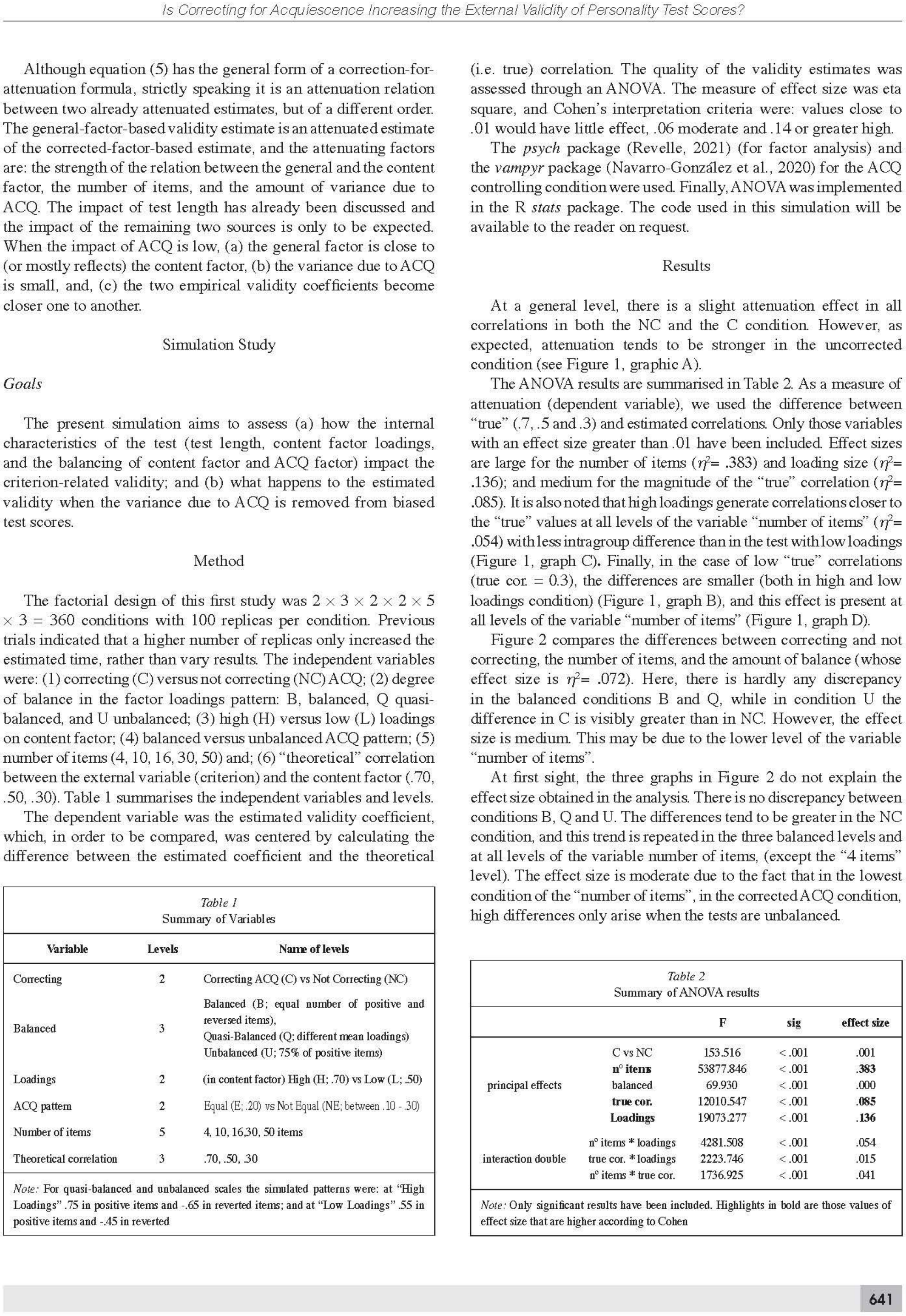
where δ_c is the covariance between the general factor scores estimates ($\hat{\theta}_g$) and the level of content (θ_c). The expression for δ_c is provided in Ferrando (2010, equation 18) and can be also estimated by regression. The main point here, however, is that, being almost a correlation (θ_c is standardized and $\hat{\theta}_g$ almost is), its value is always smaller than 1. Again, the empirical validity in (4) is an attenuated estimate of the theoretical validity. However, the attenuation is stronger here because of the additional term δ_c as well as the different amounts of information in the denominator: In effect, the error variance based on the single general factor is larger than the error variance in (3) obtained after two factors have been extracted.

The validity relation in (4) could be questioned because it is based on a wrong unidimensional model that is fitted to bidimensional data. Indeed, fitting the wrong model is expected to result in biased parameter estimates, somewhat larger item residual variances (as noted above), and, to some extent, bad model-data fit (see Ferrando, 2010). However, as long as model in (1) and (2) holds, prediction (4) continues to be correct even though the estimates provided by the unidimensional model are biased or the model does not fits well the data.

The Cost of not Correcting

The cost of not correcting for acquiescence in terms of validity, when validity is based on factor score estimates, can now be operationalized by relating the empirical validity based on general score estimates $\rho\hat{\theta}_{g'}y$ to the empirical validity based on score estimates corrected for ACQ $\rho\hat{\theta}_{c'}y$. The theoretical relation is

$$\rho\hat{\theta}_{g'}y = \rho\hat{\theta}_{c'}y\left[\delta_c \frac{\sqrt{1+\dfrac{1}{\Sigma\dfrac{\lambda_{jc}^2}{\sigma_{\varepsilon j}^2}}}}{\sqrt{1+\dfrac{1}{\Sigma\dfrac{\lambda_{jg}^2}{1-\lambda_{jg}^2}}}}\right] \tag{5}$$



Although equation (5) has the general form of a correction-for-attenuation formula, strictly speaking it is an attenuation relation between two already attenuated estimates, but of a different order. The general-factor-based validity estimate is an attenuated estimate of the corrected-factor-based estimate, and the attenuating factors are: the strength of the relation between the general and the content factor, the number of items, and the amount of variance due to ACQ. The impact of test length has already been discussed and the impact of the remaining two sources is only to be expected. When the impact of ACQ is low, (a) the general factor is close to (or mostly reflects) the content factor, (b) the variance due to ACQ is small, and, (c) the two empirical validity coefficients become closer one to another.

Simulation Study

Goals

The present simulation aims to assess (a) how the internal characteristics of the test (test length, content factor loadings, and the balancing of content factor and ACQ factor) impact the criterion-related validity; and (b) what happens to the estimated validity when the variance due to ACQ is removed from biased test scores.

Method

The factorial design of this first study was 2 × 3 × 2 × 2 × 5 × 3 = 360 conditions with 100 replicas per condition. Previous trials indicated that a higher number of replicas only increased the estimated time, rather than vary results. The independent variables were: (1) correcting (C) versus not correcting (NC) ACQ; (2) degree of balance in the factor loadings pattern: B, balanced, Q quasi-balanced, and U unbalanced; (3) high (H) versus low (L) loadings on content factor; (4) balanced versus unbalanced ACQ pattern; (5) number of items (4, 10, 16, 30, 50) and; (6) "theoretical" correlation between the external variable (criterion) and the content factor (.70, .50, .30). Table 1 summarises the independent variables and levels.

The dependent variable was the estimated validity coefficient, which, in order to be compared, was centered by calculating the difference between the estimated coefficient and the theoretical (i.e. true) correlation. The quality of the validity estimates was assessed through an ANOVA. The measure of effect size was eta square, and Cohen's interpretation criteria were: values close to .01 would have little effect, .06 moderate and .14 or greater high.

The *psych* package (Revelle, 2021) (for factor analysis) and the *vampyr* package (Navarro-González et al., 2020) for the ACQ controlling condition were used. Finally, ANOVA was implemented in the R *stats* package. The code used in this simulation will be available to the reader on request.

Table 1
Summary of Variables

Variable	Levels	Name of levels
Correcting	2	Correcting ACQ (C) vs Not Correcting (NC)
Balanced	3	Balanced (B; equal number of positive and reversed items), Quasi-Balanced (Q; different mean loadings) Unbalanced (U; 75% of positive items)
Loadings	2	(in content factor) High (H; .70) vs Low (L; .50)
ACQ pattern	2	Equal (E; .20) vs Not Equal (NE; between .10 - .30)
Number of items	5	4, 10, 16,30, 50 items
Theoretical correlation	3	.70, .50, .30

Note: For quasi-balanced and unbalanced scales the simulated patterns were: at "High Loadings" .75 in positive items and -.65 in reverted items; and at "Low Loadings" .55 in positive items and -.45 in reverted

Results

At a general level, there is a slight attenuation effect in all correlations in both the NC and the C condition. However, as expected, attenuation tends to be stronger in the uncorrected condition (see Figure 1, graphic A).

The ANOVA results are summarised in Table 2. As a measure of attenuation (dependent variable), we used the difference between "true" (.7, .5 and .3) and estimated correlations. Only those variables with an effect size greater than .01 have been included. Effect sizes are large for the number of items (η^2= .383) and loading size (η^2= .136); and medium for the magnitude of the "true" correlation (η^2= .085). It is also noted that high loadings generate correlations closer to the "true" values at all levels of the variable "number of items" (η^2= .054) with less intragroup difference than in the test with low loadings (Figure 1, graph C). Finally, in the case of low "true" correlations (true cor. = 0.3), the differences are smaller (both in high and low loadings condition) (Figure 1, graph B), and this effect is present at all levels of the variable "number of items" (Figure 1, graph D).

Figure 2 compares the differences between correcting and not correcting, the number of items, and the amount of balance (whose effect size is η^2= .072). Here, there is hardly any discrepancy in the balanced conditions B and Q, while in condition U the difference in C is visibly greater than in NC. However, the effect size is medium. This may be due to the lower level of the variable "number of items".

At first sight, the three graphs in Figure 2 do not explain the effect size obtained in the analysis. There is no discrepancy between conditions B, Q and U. The differences tend to be greater in the NC condition, and this trend is repeated in the three balanced levels and at all levels of the variable number of items, (except the "4 items" level). The effect size is moderate due to the fact that in the lowest condition of the "number of items", in the corrected ACQ condition, high differences only arise when the tests are unbalanced.

Table 2
Summary of ANOVA results

		F	sig	effect size
principal effects	C vs NC	153.516	< .001	.001
	n° items	53877.846	< .001	**.383**
	balanced	69.930	< .001	.000
	true cor.	12010.547	< .001	**.085**
	Loadings	19073.277	< .001	**.136**
interaction double	n° items * loadings	4281.508	< .001	.054
	true cor. * loadings	2223.746	< .001	.015
	n° items * true cor.	1736.925	< .001	.041

Note: Only significant results have been included. Highlights in bold are those values of effect size that are higher according to Cohen

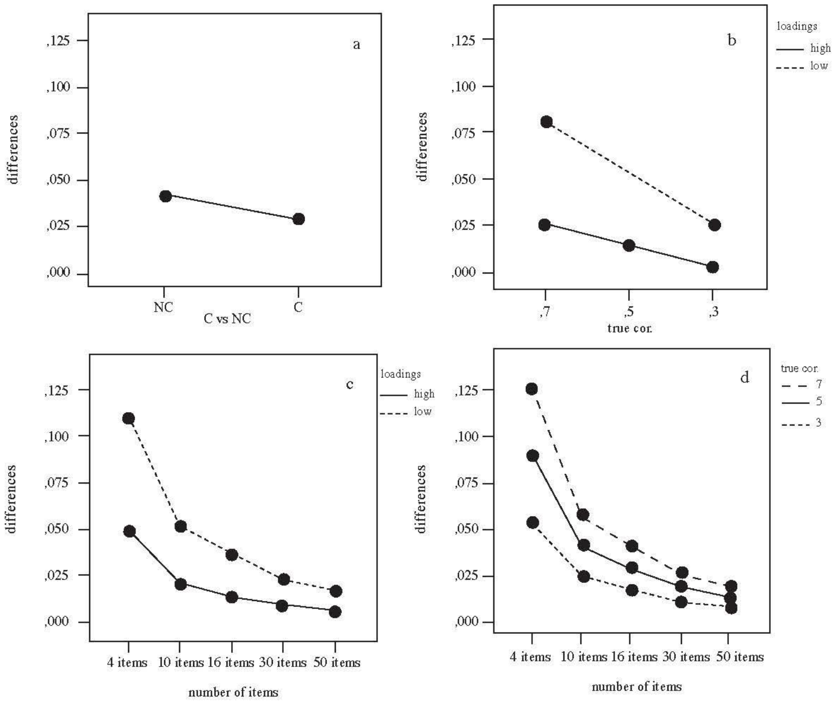


Figure 1. Graphs of ANOVA
Note: graph A: Top left; B: Top right; C: bottom left; D: bottom right

Illustrative Example 1

This first example assesses the relationship between the scores on an anxiety test, and the marks on an exam. So, the external variable can be properly considered as an objective, non-test, criterion measure.

Method

Participants

The sample consisted of 299 Psychology undergraduates (54 men and 245 women; that is, 82% of the sample were women) from six universities, with ages ranging from 18 to 60 years old (*M*=20.9 *SD*=4).

Instruments

The Examination Anxiety subscale of SAS (Statistical Anxiety Scales was used for this purpose (Vigil-Colet et al., 2008). It is a fully balanced measure made up of six 5-point Likert-type items. The numerical qualification on an exam was also recorded using a scale that ranged between 0 and 10 points (being 10 the maximum score).

Procedure

The questionnaire was administered collectively during class hours, in groups of 65 students, by a psychologist. Participation was voluntary and the protection of personal data was ensured. For one of the groups of 65 students, the teacher of the statistical subject of the group was actually the psychologist collecting the questionnaires: for this group an identification number was used so that qualifications in the subject could be added to participants' questionnaire responses. Once the data of interest was collected, any participant's identification was deleted from the study.

Data Analysis

All the computations were carried out with FACTOR (Ferrando & Lorenzo-Seva, 2017) and the Psychological Test Toolbox software (Navarro-González et al., 2019).

Results

Adequacy of correlation matrix to be factor analyzed

Data sampling adequacy was first checked and found acceptable: KMO=.810. Essential unidimensionality was assessed

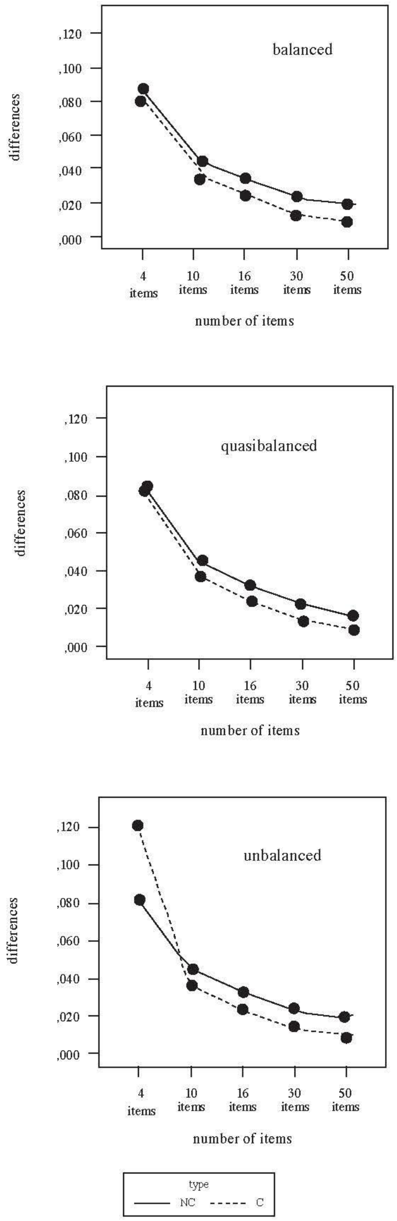


Figure 2. Triple interaction between balanced, C vs NC and number of items
Notes: graph A: balanced; B: quasi-balanced and C: unbalanced

using Explained Common Variance (ECV), and Item Residual Absolute Loadings (I-REAL) (for details, see Ferrando & Lorenzo-Seva, 2018). The ECV and I-REAL values and the corresponding 95% confidence intervals were .861 [.828, .912)], and .215 [.171, .237]. Both satisfy the thresholds of unidimensionality. Optimal Implementation of Parallel Analysis (Timmerman & Lorenzo-Seva, 2011) also recommended retaining a single factor.

Factor model fit

Robust linear exploratory factor analysis was computed using the ML criterion and specifying a single factor. Goodness of model–data fit was assessed by using both the conventional approach and the recent proposal by Yuan et al. (2017) based on equivalence testing. Goodness-of-fit measures were based on the second-order (mean and variance) corrected chi-square statistic proposed by Asparouhov & Muthen (2010) and the results are in the column labeled *Unidimensional Model* in Table 3. Overall, the results suggest that the fit of the unidimensional model is acceptable. The corresponding loading values for the unidimensional solution are in the column *Non-controlled AQ variance* (NC-ACQ) in Table 4.

Study of the impact of acquiescent variance in the factor structure

As the set of six items was fully worded balanced (i.e., half of them items are worded in the opposite direction to the other half of the items), the Ferrando et al. (2003) procedure for controlling for variance due to acquiescent responding was computed. As described in the paper by Ferrando at al, the procedure is based on an exploratory factor analytic approach: (a) it assesses the

Table 3
Goodness-of-fit values for illustrative example 1

Index	Unidimensional model	Bidimensional model
CFI	.964	.961
95% CI CFI	(.948; ,975)	(.894; .978)
T-size CFI	.919 (close)	.914 (close)
GFI	.990	.997
95% CI GFI	(.982; .997)	(.992; .999)
Z-RMSR	.069	.031
95% CI Z-RMSR	(.038; .087)	(.013; .046)

Table 4
Factor solutions for non-controlled ACQ variance (NC-ACQ) and for controlled AQ variance (C-ACQ) for illustrative example 1

Item	NC-ACQ	C-ACQ	
	EA	ACQ	EA
Direct	.916	.266	.914
Reversed	-.820	.185	-.854
Direct	.692	.373	.684
Reversed	-.422	.333	-.468
Direct	.501	.425	.503
Reversed	-.686	.315	-.779

Note: EA; Examination Anxiety

dimensionality and structure of a balanced personality scale taking into account the potential effects of acquiescent responding, and (b) it corrects the individual trait estimates for acquiescence. Goodness of fit results are in the column labeled Bidimensional Model in Table 3. Unsurprisingly, the fit of the bidimensional model is also acceptable, but is not a significant improvement on the unidimensional model. The corresponding loading matrix after controlling for ACQ is in the column *Controlled ACQ variance* (C-ACQ) in Table 4. Note that the last four items show a salient loading on the ACQ dimension (column ACQ), and that the loading of the fifth is the largest. Substantive loadings on *Examination Anxiety* when controlling for ACQ were quite similar to the ones obtained when there was no control (i.e., the unidimensional model). In fact, the value of the congruence index between the two columns is .998. The threshold value for considering two factor loading solutions to be equal is .95 (Lorenzo-Seva & ten Berge, 2006).

Validity study: Analysis of participants' acquiescent responses

EAP factor score estimates (see Ferrando & Lorenzo-Seva, 2016, for further details) on the unidimensional and the bidimensional factor solutions were computed for the 65 participants for whom qualification marks were available. For the unidimensional model, and as expected, the correlation between EA score estimates and the criterion was negative (-.368). The corresponding correlation between the "cleaned" score estimates and the criterion was -.386.

Overall, we note that the EA scale is characterized by high substantive factor weights, a relatively low decompensation between positive and negative items (see Table 4) and a small but sufficient number of items (Soto & John, 2019). In conclusion, as the predictions made here suggest, the empirical validity estimate is only slightly better when it is based on the factor score estimates corrected for ACQ.

Illustrative Study 2

In this example, we analyse the relationships between aggressiveness and intelligence, and how controlling for ACQ affects this relationship. Intelligence has often been related to violent behaviour (Ayduk et al., 2007), and this relationship has also been found using tests which control for response biases (Vigil-Colet et al., 2012; Duran-Bonavila, Morales-Vives et al., 2017). In the present study we analyse whether, as predicted by the model, the relationship increases when ACQ is controlled. It should be taken into account that all intelligence measures are maximal performance measures and, in consequence, they are not contamined by ACQ. So, we analyse the effects of removing ACQ on validity when the criterion is free from this type of bias.

Method

Participants

The sample consisted of a total of 532 students (252 men and 280 women) from 8 public high schools in the province of Tarragona, with ages ranging from 11 to 18 years old (M=14.75 SD=2.1) (see Duran-Bonavila, Vigil-Colet et al., 2017 for further details).

Instruments

The Indirect-Direct Aggression Questionnaire (IDAQ) (Ruiz-Pamies et al., 2014) provides scores for physical aggression (PA), verbal aggression (VA) and indirect aggression (IA) factors as well as an overall aggression score. Although the questionnaire has a correlated-factors structure in three dimensions, we used the overall score for two reasons. First, the tri-dimensional structure of IDAQ items depends on whether ACQ is removed or not (Navarro-González et al., 2016; Vigil-Colet et al., 2020). As a consequence, if we analyze the effects of removing acquiescence at the multidimensional level, the items comprising the solutions with and without controlling acquiescence may be different. Second, the fit of the unidimensional model after controlling for ACQ is quite acceptable (CFI=.98, RMSR=.04, RMSEA=.07), which supports the idea that the IDAQ scores measure a general factor of indirect aggression.

Three tests were used as intelligence measures: *Thurstone's Primary Mental Abilities Test* (Cordero et al., 1989), which contains scales of fluid and crystallized intelligence; *Raven's Progressive Matrices Test* (Raven, 1996), an indicator of crystallized intelligence; and *the information scale of the WAIS intelligence test for adults* (Wechsler, 2003) which is an indicator of crystallized intelligence. Intelligence measures were used as objective criterion variables, as they are all maximum performance measures and are therefore not ACQ-biased.

Procedure

School approval and parental written informed consent were obtained before participation in the study. Participation was voluntary and no incentives were given. The questionnaires were anonymous, and respondents had to provide only their gender and age.

Data Analysis

We analyzed the data reported by Duran-Bonavila, Morales-Vives et al. (2017) estimating new factor scores with and without controlling for ACQ using all the IDAQ items. Data was analyzed using the Psychological Test Toolbox (Navarro-González et al., 2019) and SPSS 25.

Results

Table 5 shows the correlations between all intelligence measures and IDAQ's overall aggression scores with and without removing ACQ effects. In all cases, the correlation between the intelligence measures and aggression is negative, a result that has consistently been obtained in previous studies (Kavish et al., 2018; González-Moraga et al., 2019). More relevant here, for all intelligence measures the correlations between aggression and intelligence were slightly larger when ACQ effects were removed. The critical threshold here appears to be -.1: when ACQ was removed, most of the correlations shown were over -.1, while they were under -.1 when no correction was used. As for differential effects, both RAVEN and WAIS had an approximate increase of .05. Correlation with RAVEN and WAIS corrected (r= -.147 and r_i= -.182) and not corrected were respectively (r = -.096 and r= -.135).

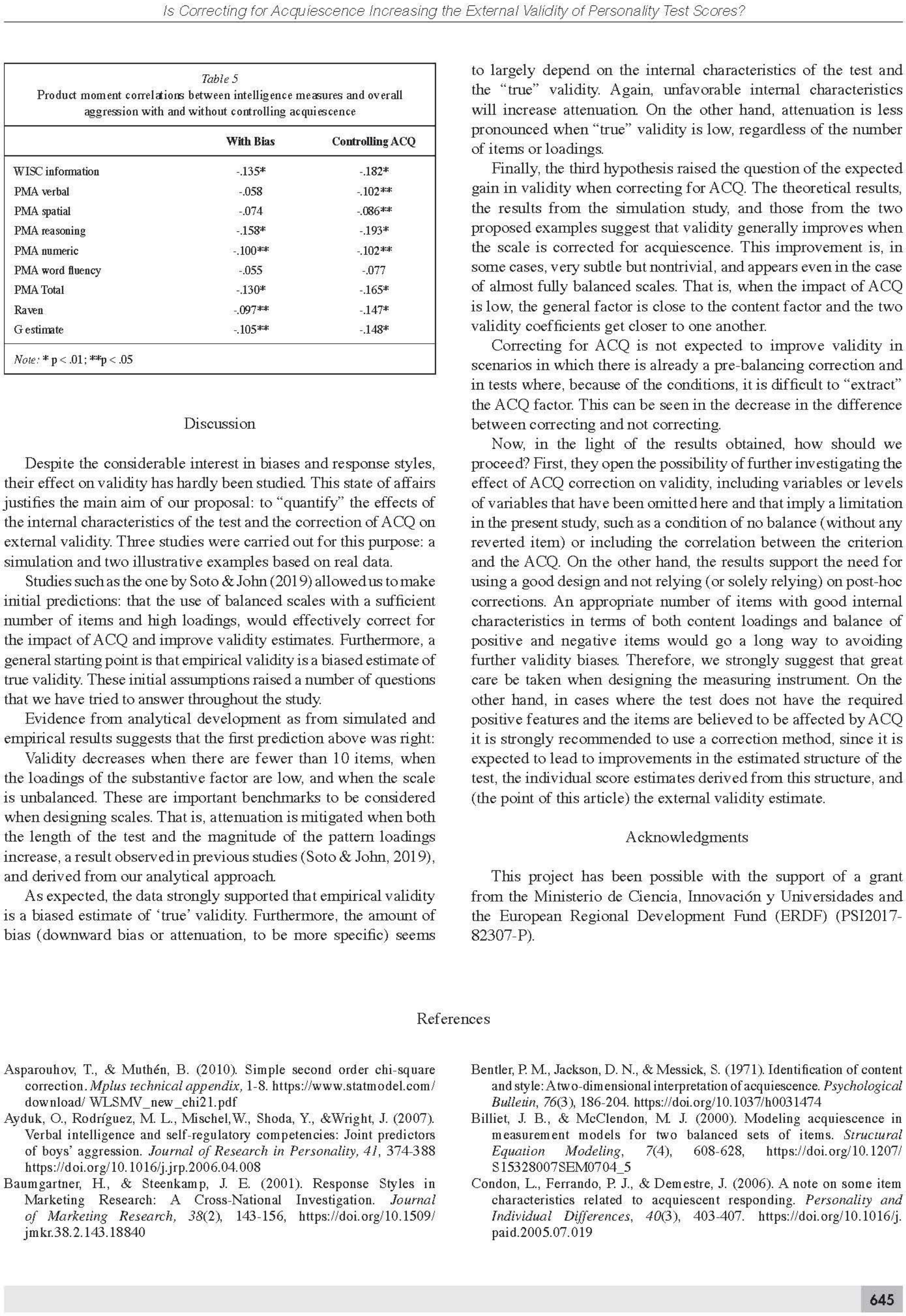


Table 5
Product moment correlations between intelligence measures and overall aggression with and without controlling acquiescence

	With Bias	Controlling ACQ
WISC information	-.135*	-.182*
PMA verbal	-.058	-.102**
PMA spatial	-.074	-.086**
PMA reasoning	-.158*	-.193*
PMA numeric	-.100**	-.102**
PMA word fluency	-.055	-.077
PMA Total	-.130*	-.165*
Raven	-.097**	-.147*
G estimate	-.105**	-.148*

Note: * p < .01; **p < .05

Discussion

Despite the considerable interest in biases and response styles, their effect on validity has hardly been studied. This state of affairs justifies the main aim of our proposal: to "quantify" the effects of the internal characteristics of the test and the correction of ACQ on external validity. Three studies were carried out for this purpose: a simulation and two illustrative examples based on real data.

Studies such as the one by Soto & John (2019) allowed us to make initial predictions: that the use of balanced scales with a sufficient number of items and high loadings, would effectively correct for the impact of ACQ and improve validity estimates. Furthermore, a general starting point is that empirical validity is a biased estimate of true validity. These initial assumptions raised a number of questions that we have tried to answer throughout the study.

Evidence from analytical development as from simulated and empirical results suggests that the first prediction above was right:

Validity decreases when there are fewer than 10 items, when the loadings of the substantive factor are low, and when the scale is unbalanced. These are important benchmarks to be considered when designing scales. That is, attenuation is mitigated when both the length of the test and the magnitude of the pattern loadings increase, a result observed in previous studies (Soto & John, 2019), and derived from our analytical approach.

As expected, the data strongly supported that empirical validity is a biased estimate of 'true' validity. Furthermore, the amount of bias (downward bias or attenuation, to be more specific) seems to largely depend on the internal characteristics of the test and the "true" validity. Again, unfavorable internal characteristics will increase attenuation. On the other hand, attenuation is less pronounced when "true" validity is low, regardless of the number of items or loadings.

Finally, the third hypothesis raised the question of the expected gain in validity when correcting for ACQ. The theoretical results, the results from the simulation study, and those from the two proposed examples suggest that validity generally improves when the scale is corrected for acquiescence. This improvement is, in some cases, very subtle but nontrivial, and appears even in the case of almost fully balanced scales. That is, when the impact of ACQ is low, the general factor is close to the content factor and the two validity coefficients get closer to one another.

Correcting for ACQ is not expected to improve validity in scenarios in which there is already a pre-balancing correction and in tests where, because of the conditions, it is difficult to "extract" the ACQ factor. This can be seen in the decrease in the difference between correcting and not correcting.

Now, in the light of the results obtained, how should we proceed? First, they open the possibility of further investigating the effect of ACQ correction on validity, including variables or levels of variables that have been omitted here and that imply a limitation in the present study, such as a condition of no balance (without any reverted item) or including the correlation between the criterion and the ACQ. On the other hand, the results support the need for using a good design and not relying (or solely relying) on post-hoc corrections. An appropriate number of items with good internal characteristics in terms of both content loadings and balance of positive and negative items would go a long way to avoiding further validity biases. Therefore, we strongly suggest that great care be taken when designing the measuring instrument. On the other hand, in cases where the test does not have the required positive features and the items are believed to be affected by ACQ it is strongly recommended to use a correction method, since it is expected to lead to improvements in the estimated structure of the test, the individual score estimates derived from this structure, and (the point of this article) the external validity estimate.

Acknowledgments

This project has been possible with the support of a grant from the Ministerio de Ciencia, Innovación y Universidades and the European Regional Development Fund (ERDF) (PSI2017-82307-P).

References

Asparouhov, T., & Muthén, B. (2010). Simple second order chi-square correction. *Mplus technical appendix,* 1-8. https://www.statmodel.com/download/ WLSMV_new_chi21.pdf

Ayduk, O., Rodríguez, M. L., Mischel,W., Shoda, Y., &Wright, J. (2007). Verbal intelligence and self-regulatory competencies: Joint predictors of boys' aggression. *Journal of Research in Personality, 41*, 374-388 https://doi.org/10.1016/j.jrp.2006.04.008

Baumgartner, H., & Steenkamp, J. E. (2001). Response Styles in Marketing Research: A Cross-National Investigation. *Journal of Marketing Research, 38*(2), 143-156, https://doi.org/10.1509/jmkr.38.2.143.18840

Bentler, P. M., Jackson, D. N., & Messick, S. (1971). Identification of content and style: A two-dimensional interpretation of acquiescence. *Psychological Bulletin, 76*(3), 186-204. https://doi.org/10.1037/h0031474

Billiet, J. B., & McClendon, M. J. (2000). Modeling acquiescence in measurement models for two balanced sets of items. *Structural Equation Modeling, 7*(4), 608-628, https://doi.org/10.1207/S15328007SEM0704_5

Condon, L., Ferrando, P. J., & Demestre, J. (2006). A note on some item characteristics related to acquiescent responding. *Personality and Individual Differences, 40*(3), 403-407. https://doi.org/10.1016/j.paid.2005.07.019

Cordero, A., Seisdedos, N., González, M., & de la Cruz, V. (1989). *PMA. Aptitudes Primarias Mentales* [Primary Mental Abilities]. TEA Ediciones.

Danner, D., Aichholzer, J., & Rammstedt, B. (2015). Acquiescence in personality questionnaires: Relevance, domain specificity, and stability. *Journal of Research in Personality, 57*, 119-130, https://doi.org/10.1016/j.jrp.2015.05.004

de la Fuente, J., & Abad, F. J. (2020). Comparing Methods for Modeling Acquiescence in Multidimensional Partially Balanced Scales. *Psicothema, 32*(4), 590-597. http://10.7334/psicothema2020.96

Duran-Bonavila, S., Morales-Vives, F., Cosi, S., & Vigil-Colet, A. (2017). How impulsivity and intelligence are related to different forms of aggression. *Personality and Individual Differences, 117*, 66-70. https://doi.org/10.1016/j.paid.2017.05.033

Duran-Bonavila, S., Vigil-Colet, A., Cosi, S., & Morales-Vives, F. (2017). How Individual and Contextual Factors Affects Antisocial and Delinquent Behaviors: A Comparison between Young Offenders, Adolescents at Risk of Social Exclusion, and a Community Sample. *Frontiers in Psychology, 8*, 1-12, https://doi.org/10.3389/fpsyg.2017.01825

Ferrando, P. J., & Lorenzo-Seva, U. (2010). Acquiescence as a source of bias and model and person misfit: A theoretical and empirical analysis. *British Journal of Mathematical and Statistical Psychology, 63*(2), 427-448. https://doi.org/10.1348/000711009X470740

Ferrando, P. J., & Lorenzo-Seva, U. (2016). A note on improving EAP trait estimation in oblique factor-analytic and item response theory models. *Psicológica, 37*(2), 235-247. https://www.redalyc.org/pdf/169/16946248007.pdf

Ferrando, P. J., & Lorenzo-Seva, U. (2017). Program FACTOR at 10: Origins, development and future directions. *Psicothema, 29*(2), 236-240 https://doi.org/ 10.7334/psicothema2016.304

Ferrando, P. J., & Lorenzo-Seva, U. (2018). Assessing the quality and appropriateness of factor solutions and factor score estimates in exploratory item factor analysis. *Educational and Psychological Measurement, 78*(5), 762-780 https://doi.org/10.1177/0013164417719308

Ferrando, P. J., Lorenzo-Seva, U., & Chico, E. (2003). Unrestricted factor analytic procedures for assessing acquiescent responding in balanced, theoretically unidimensional personality scales. *Multivariate Behavioral Research, 38*(3), 353-374, https://doi.org/10.1207/S15327906MBR3803_04

Ferrando, P. J., Morales-Vives, F., & Lorenzo-Seva, U. (2016). Assessing and controlling acquiescent responding when acquiescence and content are related: A comprehensive factor-analytic approach. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(5), 713-725. https://doi.org/10.1080/10705511.2016.1185723

González Moraga, F. R., García, D., Billstedt, E., & Wallinius, M. (2019). Facets of Psychopathy, Intelligence and Aggressive Antisocial Behaviors in Young Violent Offenders. *Frontiers in Psychology, 10*, 984. https://doi.org/10.3389/fpsyg.2019.00984

Kavish, N., Bailey, C., Sharp, C., & Venta, A. (2018). On the relation between general intelligence and psychopathic traits: An examination of inpatient adolescents. *Child Psychiatry & Human Development, 49*(3), 341-351. https://doi.org/10.1007/s10578-017-0754-8

Lechner, C. M., Partsch, M. V., Danner, D., & Rammstedt, B. (2019). Individual, situational, and cultural correlates of acquiescent responding: Towards a unified conceptual framework. *British Journal of Mathematical and Statistical Psychology, 72*(3), 426-446. https://doi.org/10.1111/bmsp.12164

Lord, F. M., & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores, Reading*. Addison-Wesley.

Lorenzo-Seva, U., & Ten Berge, J. M. (2006). Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology, 2*(2), 57-64. https://doi.org/10.1027/1614-2241.2.2.57

Lorenzo-Seva, U., & Ferrando, P. J. (2009). Acquiescent responding in partially balanced multidimensional scales. *British Journal of Mathematical and Statistical Psychology, 62*(2), 319-326. https://doi.org/10.1348/000711007X265164

Morales-Vives, F., Lorenzo-Seva, U., & Vigil-Colet, A. (2017). Cómo afectan los sesgos de respuesta a la estructura factorial de los tests basados en el modelo de los Cinco Grandes factores de personalidad [How response biases affect the factor structure of Big Five personality questionnaires]. *Anales de Psicología/Annals of Psychology, 33*(3), 589-596. https://doi.org/10.6018/analesps.33.3.254841

Navarro-González, D., Lorenzo-Seva, U., & Vigil-Colet, A. (2016). How response bias affects the factorial structure of personality self-reports. *Psicothema, 28*(4), 465-470. https://doi.org/10.7334/psicothema2016.113

Navarro-González, D., Vigil-Colet, A., Ferrando, P. J., & Lorenzo-Seva, U. (2019). Psychological Test Toolbox: A New Tool to Compute Factor Analysis Controlling Response Bias. *Journal of Statistical Software, 91*(6), 1-21. https://doi.org/10.18637/jss.v091.i06

Navarro-González, D., Vigil-Colet, A., Ferrando, P.J., Lorenzo-Seva, U., & Tendeiro, J.N. (2020). *vampyr: Factor Analysis Controlling the Effects of Response Bias* (version 1.1.1) [R package]. https://cran.rstudio.com/web/packages/vampyr/index.html

Primi, R., De Fruyt, F., Santos, D., Antonoplis, S., & John O. P. (2019). True or False? Keying Direction and Acquiescence Influence the Validity of Socio-Emotional Skills Items in Predicting High School Achievement. *International Journal of Testing 20*(2), 97-121. https://doi.org/10.1080/15305058.2019.1673398

Primi, R., Santos, D., De Fruyt, F., & John, O. P. (2019). Comparison of classical and modern methods for measuring and correcting for acquiescence. *British Journal of Mathematical and Statistical Psychology, 72*(3), 447-465. https://doi.org/10.1111/bmsp.12168

R Core Team (2013). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. http://www.R-project.org/

Raven, J. C. (1996). *Matrices progresivas. Escalas CPM Color y SPM General* [Raven Progressive Matrices]. TEA Ediciones.

Ray, J. J. (1983). Reviving the problem of acquiescent response bias. *The Journal of Social Psychology, 121*(1), 81-96, http://doi.org/10.1080/00224545.1983.9924470

Revelle, W. (2021). *Psych: Procedures for psychological, psychometric, and personality research* (version 2.1.6) [R package]. https://cran.rstudio.org/web/packages/psych/psych.pdf

Ruiz-Pamies, M., Lorenzo-Seva, U., Morales-Vives, F., Cosi, S., & Vigil-Colet, A. (2014). I-DAQ: A new test to assess direct and indirect aggression free of response bias. *The Spanish Journal of Psychology, 17*, E41. https://doi.org/10.1017/sjp.2014.43

Soto, C. J., & John, O. P. (2019). Optimizing the length, width, and balance of a personality scale: How do internal characteristics affect external validity? *Psychological Assessment, 31*(4), 444-459. https://doi.org/10.1037/pas0000586

Soto, C. J., John, O. P., Gosling, S. D., & Potter, J. (2008). The developmental psychometrics of big five self-reports: Acquiescence, factor structure, coherence, and differentiation from ages 10 to 20. *Journal of Personality and Social Psychology, 94*(4), 718-737. https://doi.org/10.1037/0022-3514.94.4.718

Suárez Álvarez, J., Pedrosa, I., Lozano, L. M., García Cueto, E., Cuesta Izquierdo, M., & Muñiz Fernández, J. (2018). Using reversed items in Likert scales: A questionable practice. *Psicothema, 30*(2), 149-158. http://10.7334/psicothema2018.33

Timmerman, M. E., & Lorenzo-Seva, U. (2011). Dimensionality assessment of ordered polytomous items with parallel analysis. *Psychological Methods, 16*(2), 209-220. https://doi.org/10.1037/a0023353

Vigil-Colet, A., Lorenzo-Seva, U., & Condon, L. (2008). Development and validation of the statistical anxiety scale. *Psicothema, 20*(1), 174-180. http://www.psicothema.com/pdf/3444.pdf

Vigil-Colet, A., Navarro-González, D., & Morales-Vives, F. (2020). To reverse or to not reverse Likert-type items: That is the question. *Psicothema, 32*(1), 108-114. https://doi.org/10.7334/psicothema2019.286

Vigil-Colet, A., Ruiz-Pamies, M., Anguiano-Carrasco, C., & Lorenzo-Seva, U. (2012). The impact of social desirability on psychometric measures of aggression. *Psicothema, 24*(2), 310-315. https://www.redalyc.org/pdf/727/72723578021.pdf

Wechsler, D. (2003). *Escala de inteligencia de Wechsler para niños-IV (WISC-IV)* [Wechsler Intelligence Scale for Children-WISC-IV]. Psychological Corporation.

Wetzel, E., Böhnke, J. R., & Brown, A., (2016). Response biases. In F. T. L. Leong, D. Bartram, F. Cheung, K. F. Geisinger, & D. Iliescu (Eds.), *The ITC International Handbook of Testing and Assessment* (pp. 349-363). Oxford University Press.

Yuan, K. H., Chan, W., Marcoulides, G. A., & Bentler, P. M. (2016). Assessing structural equation models by equivalence testing with adjusted fit indexes. *Structural Equation Modeling, 23*(3), 319-330. https://doi.org/10.1080/10705511.2015.1065414

SIREN: An Hybrid CFA-EFA R Package for Controlling Acquiescence in Restricted Factorial

by David Navarro-Gonzalez, Pere J. Ferrando, Fabia Morales-Vives, Ana Hernandez-Dorado

Abstract The **siren** package implements a two-step procedure that allows restricted (confirmatory) factor analytic (FA) solutions to be fitted in data matrices that have been previously 'cleaned' of the biasing effects of acquiescent responding (AR) by using an unrestricted (exploratory) FA specification. So, the procedure which is implemented is hybrid: i.e. (a) an unrestricted acquiescence (ACQ) factor is first fitted to the data, (b) the residual data (or covariance) matrix after the impact of ACQ has been partialled-out is obtained, and (c) a restricted FA solution is fitted to the residual matrix. Although the basic foundations of the procedure are known, it contains new methodological developments that are, all of them, implemented in the package. So, provided that fully or partially balanced scales are available, the researcher will be able to: (a) calibrate a multidimensional CFA solution which is free from AR, (b) assess the goodness of model-data fit of this solution, and (c) obtain individual score estimates in the content as well as in the ACQ factors. The functioning of the program is assessed by means of a simulation study, and illustrated with a toy example. Its usefulness is also demonstrated by using an illustrative example in the personality domain. **siren** is submitted to be a valuable tool for use in item CFA applications when AR is expected to be operating.

1 Introduction

Valid interpretation of typical-response or non-cognitive (personality, attitude, interest, etc.) test scores require that the data used in the calibration of the test items meet a series of conditions. Of these, one of the more basic is that the responses truly reflect the influence of the content variables intended to be measured, and are not affected by other systematic determinants, generally unrelated to content. Among these unwanted determinants, this article is concerned with Acquiescent Responding (AR): the tendency to agree or endorse an item regardless of its content (Messick, 1966). Tests designers and practitioners are generally aware of the potential invalidating effects of AR, and use procedures for controlling them. Of these, the most common is to use fully or partially balanced scales, which are made of items keyed in opposite directions of the content variables (Savalei and Falk, 2014; Vigil-Colet et al., 2020). Statistical control of AR in balanced scales is generally based on factor analytic (FA) procedures, and essentially entails explicitly modeling AR as an additional factor in order to avoid the biasing effects that AR would have if left unmodeled. This FA-based control operates at two levels: first, at the level of the factor structure obtained in the calibration stage; and second, at the level of the factor score estimates derived from the calibration structure (Ferrando et al., 2003).

Within the general FA modeling, two main approaches exist at present (Savalei and Falk, 2014; de la Fuente and Abad, 2020). The first is fully confirmatory, and the solution is identified by restricting all the loadings on the additional Acquiescence (ACQ) factor to have the same unit value (Billiet and McClendon, 2000). The second is exploratory or semi-confirmatory ((Ferrando et al., 2003)): First, (a) an unrestricted ACQ factor with (possibly) different loadings and (b) an also unrestricted (EFA) direct "content" solution are obtained. Second, the direct content solution is either analytically rotated (fully exploratory solution) or rotated against a specified or semi-specified target (semi-confirmatory solution). The pros and cons of both approaches have been discussed and compared by Savalei and Falk (2014) and de la Fuente and Abad (2020). Both studies concluded that the confirmatory approach is more robust and user-friendly than the EFA with target rotation. However, it is also more sensitive to violation of the unit-weight loading assumption for the ACQ factor. The aim of this paper is to propose an implement a "hybrid" approach, named SIREN, that combines CFA and EFA features and that, furthermore, is (a) intended for fitting multiple content solutions and (b) based on scales that will not generally be fully balanced. Because we are using the same name for the proposed procedure and the package that implements it, in the remaining of the paper, we shall use the distinction "SIREN procedure" and "**siren** package" when necessary so as to avoid confusion.

At the calibration level, the basic idea of the SIREN procedure is to first obtain an ACQ factor in which the loadings are not restricted to have the same value. Next, the impact of this factor is partialled out, and finally a restricted or confirmatory FA solution is fitted to the "cleaned" data. At the scoring level, ACQ and content individual score estimates are obtained based on the unrestricted ACQ pattern and the restricted, CFA content solution. Although most basic foundations of SIREN are known in the FA literature (e.g. Nunnally, 1978), the full proposal contains new developments, and as a whole is, we believe, a new contribution.

Preliminary Considerations

Consider a set of *n* items intended to measure p common content factors (e.g. personality dimensions). The basic FA model equation in the population is:

$$\mathbf{Z} = \boldsymbol{\Lambda}\boldsymbol{\theta} + \boldsymbol{\Psi}\mathbf{E} \tag{1}$$

where **Z** is an $n \times 1$ random vector of observed item scores; $\boldsymbol{\Lambda}$ is an $n \times p$ factor pattern matrix; $\boldsymbol{\theta}$ is an $p \times 1$ random vector of 'true' common factor scores; $\boldsymbol{\Psi}$ is an $n \times n$ diagonal matrix of unique-factor loadings, and **E** is an $n \times 1$ random vector of unique factor scores. The reproduced covariance matrix among the *n* item scores as implied by model 1 is given by the structural equation:

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}^2 \tag{2}$$

where $\boldsymbol{\Phi}$ is $p \times p$ correlation matrix containing the correlations between the 'true' common factor scores. Generally, in the applications considered here, the **Z** scores will be standardized scores, and so, the implied covariance matrix $\boldsymbol{\Sigma}$ in 2 will be a correlation matrix.

The main difference between an unrestricted (exploratory) and a restricted (confirmatory) solution within the general model 2 is in the constraints that are imposed to the pattern matrix $\boldsymbol{\Lambda}$. In an unrestricted solution, only minimal identification constraints are imposed, so that the common space: $\boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}'$ in 2 is not restricted and multiple solutions of the same type, that fit all equally well, can be obtained from each other by rotation. In a restricted solution, the number of imposed restrictions makes the specified solution $\boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}'$ unique, in the sense that it cannot be obtained by rotation of another solution (see Joreskog). Although a restricted solution can be obtained by using different sets of constraints, the most usual consist of imposing an independent-cluster structure (e.g. McDonald, 2000): each item has only a non-zero loading in one factor, having zero loadings in all the others.

At this point, we will start to develop a small, artificial toy example to help clarify the explanations that will follow. Suppose a questionnaire made up of 8 factorially simple items that measure two moderately correlated factors, so that the independent-cluster structure in the population is:

Table 1: Toy example: restricted solution with two correlated content factors.

$$\boldsymbol{\Lambda} = \begin{bmatrix} .7 & 0 \\ -.7 & 0 \\ .7 & 0 \\ -.7 & 0 \\ 0 & .6 \\ 0 & -.6 \\ 0 & .6 \\ 0 & -.6 \end{bmatrix} \quad \boldsymbol{\Phi} = \begin{bmatrix} 1 & .3 \\ .3 & 1 \end{bmatrix}$$

A CFA estimation of this structure can be specified by constraining to zero the 8 elements of $\boldsymbol{\Lambda}$ that should be zero and freely estimating the remaining 8 loadings and the interfactor correlation based on the sample correlation matrix **R**. Note that, for a solution of this type to be defined, the practitioner must be able to specify: (a) the number of content factors that the questionnaire intends to measure (two in the example), and (b) the specific items that define each factor. Furthermore, the items are supposed to be all factorially simple, so that each item is a marker of the factor it measures and has negligible loadings on the remaining factors. These conditions are not easy to achieve, but can be feasible at advanced stages of test development.

Suppose now that the **content** structure of our example is that in 1 but, at the same time, the item responses are also partly affected by AR, conceptualized as an additional non-content factor (see 2 below). Now, even though the content structure was correct, if the specified two-factor structure above was directly fitted to R, the expected results would be that (a) the goodness of model-data fit would not be good, and (b) the loading and inter-factor correlation estimates would be biased with respect to the parameter values in (2) (see e.g. DeMars, 2014; Ferrando and Lorenzo-Seva, 2010).

At this point, the rationale of SIREN becomes clear. It is a matter of obtaining a corrected or cleaned covariance or correlation matrix $\mathbf{R}_{\text{corr}}$ in which the impact of the ACQ factor has been partialled-out. If this is done correctly, and the specified solution is fitted to $\mathbf{R}_{\text{corr}}$ instead of **R**, the solution will now fit well, and the 'true' content parameters in 1 will be well recovered. And not only this, improved individual content scores that are free of the biasing effects of AR, and individual estimates of the ACQ levels will also be obtained for each respondent.

As mentioned above, in order to control for the impact of AR, the items of the questionnaire have to be fully or partially balanced. In the present scenario, the condition of full balance implies that, within each factor, half of the items that define this factor are positively keyed and the other half are negatively keyed. The condition of partial balance implies here that all the factors contain positively and negatively keyed items, but that the number of positive and negative items is not the same at least in one factor (e.g. Lorenzo-Seva and Ferrando, 2009)

We shall now illustrate the points so far discussed with our toy example. Suppose now that the full content plus ACQ structure in the population is that in 2. As the content pattern loadings show, however, the practitioner, has done her work well and, within each factor the items are fully balanced: within each content factor half of the loadings are positive and half negative. As for the ACQ factor, (a) all the loadings are positive, and (b) they are smaller in magnitude than the content loadings. Both features are expected in empirical applications. First, AR is the tendency to agree with the item regardless of the direction in content (hence all loadings are expected to be positive). Second, in a well-designed measure, the item responses are expected to be far more determined by the content they measure than by ACQ.

Table 2: Toy example: complete solution when ACQ is operating. Balanced items.

$$\mathbf{\Lambda} = \begin{bmatrix} .7 & 0 & .1 \\ -.7 & 0 & .2 \\ .7 & 0 & .3 \\ -.7 & 0 & .3 \\ 0 & .6 & .3 \\ 0 & -.6 & .3 \\ 0 & .6 & .2 \\ 0 & -.6 & .1 \end{bmatrix} \qquad \mathbf{\Phi} = \begin{bmatrix} 1 & .3 & 0 \\ .3 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Description of the procedure and relation with previous approaches

We propose a multi-stage procedure in which the number of stages depends on whether the test is fully or only partially balanced. So, the stages will be described separately for the two scenarios. Conceptually, however, it is useful to view the overall procedure as based on three general stages. In the first stage, an ACQ factor is estimated from the properties of the (partially or fully) balanced set of items, and the impact of this factor on the inter-item correlation matrix is partialled-out. In the second stage, a specified CFA solution is fitted to the 'cleaned' correlation matrix. Finally, in the third stage, individual score estimates are obtained from the hybrid solution (i.e. the unrestricted ACQ factor and the restricted content-factor solution).

Although the sequential rationale just described is conceptually the clearest, the structural solution at the second stage above can actually be specified and fitted in two ways. The first way directly follows from the corrected-correlation-matrix concept: to fit a CFA solution to a reduced correlation matrix which is free from ACQ. The second way is to take the estimated ACQ loadings obtained at the first stage as if they were fixed and know, and next to specify a full CFA solution that includes an additional ACQ factor with loadings fixed at the obtained values. As we shall see, the results from both approaches must be the same.

The explanation above allows the relation between SIREN and previous approaches to be discussed in more detail. Initially SIREN was developed as a hybrid extension of Ferrando et al. (2003)'s EFA approach. Instead of obtaining a fully (i.e. content plus ACQ) unrestricted solution that is next jointly rotated, as originally proposed, only the ACQ factor is obtained here first in an unrestricted way, and next a CFA solution is fitted to the residual matrix. Conceptualized in this way, SIREN can be viewed as a particular application of a residual covariance analysis approach, i.e. to initially correct a covariance or a correlation matrix for unwanted effects before it is used as input for further structural analyses (e.g. Andrews, 1984; Asparouhov and Muthen, 1984; DeCastellarnau and Saris, 2021; ten Berge, 2020)

If the equivalent specification of (a) taking the first-stage estimated ACQ loadings as if they were fixed and know, and (b) specifying a full CFA solution that includes the additional fixed ACQ factor, is used instead, then SIREN can be regarded as a modification of the CFA approach initially proposed by Billiet and McClendon (2000). In effect, in the latter, the ACQ loadings are fixed all of them to unity for identification purposes (which is generally unrealistic). In contrast, in SIREN these loadings are fixed at the (possibly different) values estimated at the first stage.

Basic Results

Consider again a fully-balanced questionnaire made up of n items (where n is even) that measure a set of (possibly related) traits $\theta_1 ... \theta_l ... \theta_m$, so that each item is a factorially pure measure of one of the m content factors plus of an acquiescence factor θ_a a which is unrelated to the content factors. For an individual i that responds to an item j that measures content factor l, the structural model in a z-score metric (mean 0 and variance 1) is

$$z_{ij} = \lambda_{jl}\theta_{il} + \alpha_{ja}\theta_{ia} + \varepsilon_{ij} \tag{3}$$

This is a scalar specification of matrix equation 1 based on an independent-cluster pattern. Only a single content loading per item is specified because the remaining content loadings are zero. The α_{ja} loading is the loading item j has on the ACQ factor. Finally, the residual terms εs have zero means, and are uncorrelated with the factors or with one another.

The z item responses in 3 can be treated as categorical or (approximately) continuous. In the first case, the standardized scores would correspond to the strength response variables that underlie the observed responses (see Muthen, 1993). In the second case, they are directly the standardized item scores. From this general modeling, it follows that the inter-item correlations are polychoric correlations in the first case, and product-moment correlations in the second case (see e.g. Ferrando and Lorenzo-Seva, 2013, for further details). The correlational results that follow are common for both treatments.

Consider now the (polychoric or product-moment) reduced inter-item correlation matrix with communalities in the main diagonal (see Ferrando and Lorenzo-Seva, 2010). If (a) all the assumptions so far (independent-cluster structure, full balance within factors) were met, (b) the specified FA model was correct, and (c) the item communalities were known, then the first centroid loading (e.g. Lawley, 1960) for item j would correspond to the loading this item has on the ACQ factor (i.e. α_{ja}) in the population (see e.g. Ferrando and Lorenzo-Seva, 2010, for further details). Indeed, the conditions above are only approximately met at best (in particular, true communalities are never known; (McDonald, 1978, see). And, furthermore, the first centroid loading is a sample estimate. For these reasons, our choice in SIREN is to (a) obtain the first principal-axis or canonical factor using a modern and efficient EFA estimation procedure, and (b) rotate this factor against the centroid vector that is used as a target or criterion (see Eysenck, 1950). In general, the first canonical factor and the first target centroid are already very close (Chulakian, 2003; ten Berge, 2020), and so, the modifications due to the target rotation are small. The final factor so obtained is taken in SIREN as an estimate of the ACQ factor.

With regards to the choice of the more efficient EFA procedure, SIREN uses Minimum Rank Factor Analysis (MRFA; ten Berge and Kiers, 1991), a quite robust unweighted least squares (ULS) procedure, that works particularly well when weak common factors are expected and the sample is not too large. Furthermore, MRFA provides estimates of the proportion of common variance accounted for by the different factors, and this information is useful for assessing the relevance of the ACQ factor in terms of explained common variance.

Multi-stage approach with fully balanced scales

Stage 1: for the n test items, the ACQ factor loadings are estimated by: (a) obtaining the first MRFA factor of the inter-item correlation matrix, (b) obtaining the first centroid of this matrix according to equation 4 (to be used as a target), and (c) rotating the MRFA factor to the position of maximal congruence with respect to target (b). Denote by $\mathbf{R}$ the inter-item correlation matrix, and by $\boldsymbol{\alpha}$ the column of estimated loadings obtained at the end of step (c). Stage 2: obtain the corrected (i.e. ACQ free) inter-item residual matrix as $\mathbf{R}_{\mathbf{corr}} = \mathbf{R} - \boldsymbol{\alpha}\boldsymbol{\alpha}'$. The $\mathbf{R}_{\mathbf{corr}}$ matrix is, and should be treated as, a residual covariance matrix (not a correlation matrix). Stage 3: The prescribed CFA solution can be specified and fitted in two alternative ways. The first is to input $\mathbf{R}_{\mathbf{corr}}$ specified as a covariance matrix to the SEM program, and request a standardized solution. The output will consist of the standardized content pattern with loadings that are free of ACQ. The second way is to input the raw data to the SEM program, by specifying the prescribed CFA content solution, plus an additional ACQ factor in which all the loadings are specified as fixed and known. In this second specification, the ACQ loadings $\boldsymbol{\alpha}$ obtained in stage 2 cannot be imputed directly because $\mathbf{R}_{\mathbf{corr}}$ is a covariance matrix (i.e. unstandardized) while the loadings on $\boldsymbol{\alpha}$ are standardized. The correct values to be imputed are thus those obtained by multiplying each standardized loading on $\boldsymbol{\alpha}$ by the corresponding item standard deviation: $\boldsymbol{\alpha}_x$(scaled)= $\boldsymbol{\alpha}_x \mathbf{sx}$. This scaling transforms the standardized loadings $\boldsymbol{\alpha}$ into unstandardized loadings (e.g. Bollen, 1989). As in the first approach, a standardized solution will next be requested in the output. If this is done, the output will now consist of (a) the standardized content pattern with loadings free of ACQ and (b) an additional column containing the standardized ACQ loadings. Indeed, the standardized content pattern must be the same in both specifications.

The CFA in stage 3 is done by using the cfa function from the **lavaan** and one of the two options described above. However, of the multiple choices that the program allows for estimating the structural item content parameters, we have chosen the one that is most congruent with the previous stages. Thus, whether the variables are treated as continuous or discrete, the estimation procedure is robust ULS, in agreement with the choice of MRFA-ULS for obtaining the ACQ estimates.

Stage 4: testing model-data fit at the structural level. Again, we have made choices that are in accordance with the limited-information nature of the procedure and with the results of the simulation study (see below). The chosen indices are: (a) the RMSR and GFI as overall measures of misfit (McDonald and Mok, 1995), (b) the RMSEA as a measure of relative fit with respect to the degrees of freedom (i.e. model complexity), and (c) the CFI as a measure of comparative fit with respect to the null independence model (see Tanaka, 1993, for points b and c).

Stage 5: obtaining individual score estimates. In the basic FA equation discussed above, the factor score estimates for each individual are the estimates of the 'true' scores $\boldsymbol{\theta}$ in 1, which, of course, are unknown. For both, the linear and the nonlinear models, the factor score estimates are Bayes Modal a Posteriori (MAP), which, in the continuous (linear) model, are known as regression estimates. In both cases, Bayesian scoring provides finite and plausible estimates for all the respondents under study (see Ferrando and Lorenzo-Seva, 2016). For each participant, the output information consists of the point estimate of his/her level on the content factors plus his/her factor score estimate on the ACQ factor. This last estimate can be interpreted as the predisposition of the individual to engage in AR.

Stages 1 to 4 in the fully-balanced procedure will be now illustrated with our toy example. Furthermore, the effects of ignoring the secondary ACQ factor will be illustrated by fitting directly the content solution in table 1 to the uncorrected (i.e observed) correlation matrix. To perform the illustration, we generated a random sample of $N = 500$ simulees from a population in which the complete solution in 2 holds. The results are in 3.

Table 3: Toy Example Results

Siren Loadings	Direct Loadings
$\boldsymbol{\Lambda} = \begin{bmatrix} .758 & 0 & .118 \\ -.664 & 0 & .175 \\ .62 & 0 & .311 \\ -.723 & 0 & .257 \\ 0 & .605 & .382 \\ 0 & -.583 & .272 \\ 0 & .681 & .137 \\ 0 & -.582 & .165 \end{bmatrix}$	$\boldsymbol{\Lambda} = \begin{bmatrix} .761 & 0 \\ -.693 & 0 \\ .631 & 0 \\ -.727 & 0 \\ 0 & .549 \\ 0 & -.552 \\ 0 & .720 \\ 0 & -.586 \end{bmatrix}$
Siren Phi Matrix	**Direct Phi Matrix**
$\boldsymbol{\Phi} = \begin{bmatrix} 1 & .33 & 0 \\ .33 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$	$\boldsymbol{\Phi} = \begin{bmatrix} 1 & .35 \\ .35 & 1 \end{bmatrix}$
Siren GOF	**Direct GOF**

CFI = 1
GFI=.998
RMSEA=0
SRMR= .024
	CFI = .922
	GFI=.974
	RMSEA=.088
	SRMR= .049

Note: GOF=Goodness of Fit Indices

Results in 3 are predictable Ferrando and Lorenzo-Seva (2010) and DeMars (2014) can be summarized as follows. With regards to bias, SIREN does a good job, and recovers quite acceptably (given both the sample and model size) all the parameters in the toy solution: content loadings, ACQ loadings, and inter-factor correlation. On the other hand, the loading estimates in the uncontrolled solution tend to be slightly more biased, and the inter-factor correlation slightly over-estimated. In terms of GOF, that of the SIREN solution is almost perfect by all standards, which is only to be expected, as the specified solution is correct. The GOF of the uncorrected solution, however, clearly deteriorates.

The amount of misfit is not terribly bad here, which is also expected in such a small model. In a larger-sized example, however, the deterioration of fit would have been much stronger.

Multi-stage approach with partially balanced scales

The basic idea in this case is to first obtain a fully balanced sub-set of items, which we shall denote as the core set, and estimate the ACQ loadings of the items belonging to this core by using the procedure described above. Next, the ACQ loadings of the remaining items are estimated by using a type of extension analysis based on the method of moments. Once the ACQ loading estimates are available for all the test items, the rest of the procedure can be carried out exactly as in the fully balanced case. So, the points that require specific discussion are (a) how to determine which items will be included in the core set, and (b) how the ACQ loadings of the remaining items will be determined.

Stage 1: Choosing the core set. Within each specified factor, the positive and negative items are separated into two groups, and a centroid FA as in 4 is performed separately in each of the two resulting inter-item correlation matrices. The loadings in the smaller set (usually that containing the negative items) are taken as fixed, and each of them is paired with the positive loading with the most similar value. The aim is for the absolute value of the sums of the positive and negative loadings to be as similar as possible. The rationale is that the effect of ACQ will be in the same direction if items are all worded in the same direction (i. e. both the positive and the negative loadings will be upwardly biased).

Stage 2: For the n_c items in the core subset, the loadings on the ACQ factor are estimated using the procedure described in the fully balanced case.

Stage 3: Denote by X_o an item outside the core set, and let $j = 1, \cdots n_c$ be the items in the core. Let $\sum_{j=1}^{n_c} r_{oj}$ be the sum of the correlations of item X_o with the remaining items in the core set. If the core items are balanced, it then follows that:

$$\sum_{j=1}^{n_c} r_{oj} = \lambda_{oa} \sum_{j=1}^{n_c} \lambda_{ja} \tag{4}$$

So

$$\lambda_{oa} = \frac{\sum_{j=1}^{n_c} r_{oj}}{\sum_{j=1}^{n_c} \lambda_{ja}} \tag{5}$$

In words, if full balance holds for the core set, then the quotient between (a) the sum of correlations of item X_o with the remaining items in the core set, and (b) the sum of ACQ loadings in the core set provides a simple estimate of the loading of item X_o on the ACQ factor. Note that the sum in the denominator of 5 is taken as fixed and known and has been obtained in Stage 2 above. The estimate described above can be viewed as an extension-analysis estimate (e.g. McDonald, 1978) obtained by the method of moments.

The extension estimate 5 is computed on an item-by-item basis for each of the items outside the core set in Stage 3. So, at the end of this stage, ACQ loading estimates are available for all the test items under study. This is the same situation as at the end of Stage 1 in the fully-balanced-case approach. Therefore, from this point on, the procedure is the same in both cases.

2 Further mathematical and statistical details

Provided that the preliminary general conditions discussed above for using SIREN are met, the correct functioning of the procedure depends on two main points. The first point is of a mathematical nature, and is critical if unbiased loadings (especially those of content factors) are to be obtained. The second is of a statistical nature, and is important if the goodness-of-fit assessment needs to be correct.

Obtaining unbiased loading estimates in SIREN mainly depends on achieving full balance (either for the total test or for the core set), in principle, within each factor. In more detail, what is required is that the sum of content loadings be equal to zero. If it is, then the first centroid (or MRFA factor) will reflect only AR, so partializing it from the correlation matrix will remove only this response bias and leave a 'clean' corrected correlation that will reflect only content. However, if balance is not achieved, then, to a greater or lesser extent, the first centroid factor will reflect a mixture of ACQ and content. So, some content will be removed by partializing, and the resulting content loadings will be biased (ten Berge, 2020).

From the discussion above, it seems clear that SIREN is not expected to work in a purely exploratory analysis based on a set of items, of which some are balanced, although the factor they belong to is not known. In our opinion it is dubious that a proper ACQ factor could be identified in these circumstances.

Having said that, however, we also note that the strict condition of full balance within each factor (or the core set within each factor) is probably too strong. Preliminary results by the writers suggest that, provided that (a) the content CFA solution is correct, and (b) full balance holds for the entire set (or core set) of items but not necessarily for each factor, then SIREN would still perform correctly in most cases. For the moment, however, we prefer to maintain the strong within-factor balance requirement to guarantee the proper functioning of the method. The extent to which the content loading estimates will degrade as imbalance increases is best addressed using simulation, and this will be done in the next section.

We turn now to the second, statistical point. Our proposal can be viewed as a particular application of what Nunnally (1978) called an ad-lib factorial process. Nunnally considered that it was entirely legitimate to fit successive factors to residual matrices by using different methods. So, once a residual matrix had been obtained, it could be fitted by any method regardless of the one that had been used to obtain the residual. While we agree that our proposal is indeed legitimate, its multi-stage nature necessarily entails a loss of information.

Consider first the fully balanced case. The ACQ estimates are first obtained by using MRFA, essentially a ULS procedure. A restricted CFA solution is then fitted to the resulting residual matrix, and this matrix is fitted by the robust ULS procedures mentioned above. However, the CFA estimates obtained are, in fact, conditional upon the MRFA-ACQ estimates obtained previously. These estimates are taken as fixed and known, and their uncertainty is not taken into account. So, the estimator used to fit the content solution is less efficient than if it had been based on a covariance or correlation matrix directly obtained from the observed data. In the partially balanced case, the loss of information-efficiency is more marked: the core-set of ACQ loading estimates are MRFA estimates. The extension ACQ estimates are moment estimates conditional upon the core estimates. And, finally, the content CFA estimates are conditional upon all the ACQ estimates previously obtained.

In agreement with authors such as: DeCastellarnau and Saris (2021); Nunnally (1978); ten Berge (2020), and the empirical results provided by Oberski and Satorra (2013), we believe that the impact of the loss of efficiency discussed above on the point estimates and indices of goodness of fit will be relatively minor in practice provided that the proposed solution is correct and the basis conditions are reasonably met. The issue, however, needs to be, and will be, assessed by using simulation. In any case, we regard the use of SIREN as a trade-off, and believe that the impact of model misspecification (i.e. fitting a content solution that ignores ACQ when ACQ is in fact operating) is far worse in terms of biases and GOF results than the loss of efficiency due to an ad-lib factoring approach that is based on a more correct solution.

3 The siren package details

Available through CRAN, the **siren** package contains one main function (and additional internal functions) called `acquihybrid`, which implements the procedures described in the sections above.

The function usage is the following:

```
acquihybrid(x, content_factors, target, corr = "Pearson", raw_data=TRUE,
method="fixed", display = TRUE)
```

in which the arguments are:

`x`, raw sample scores or a covariance/correlation matrix,

`content_factors`, the number of content factors to be retained. Each factor has to be defined by 3 items,

`target`, the target matrix, which provides the signed dominant loading of each item on its corresponding factor. The target is only used as a reference for assessing which items have significant loadings on which factors, and the exact value is not used,

`corr`, determines the type of matrices to be used in the factor analysis. "Pearson": Computes Pearson correlation matrices (linear model); "Polychoric": Computes Polychoric/Tetrachoric correlation matrices (graded model),

`raw_data`, logical argument, if TRUE, the entered data will be treated as raw scores (default). If FALSE, the entered data will be treated as a covariance/correlation matrix,

`method`, two choices are provided: fixed, which use the ACQ loadings obtained in the first step to specify the ACQ factor in the CFA solution based on the direct scores, and resid, which uses the ACQ-free covariance matrix as input for the CFA,

`display`, determines if the output will be displayed in the console, TRUE by default. If it is TRUE, the output is printed in the console and if it is FALSE, the output is returned silently to the output variable.

The data provided should be a data frame or a numerical matrix for input vectors and matrices, character variables for `corr` and method arguments, and logical values for `raw_data` and `display` arguments.

The `acquihybrid` function returns a list variable, containing the following variables:

`rloadings`, the factor loadings for each content factor and acquiescence factor.

`rfactor_cor`, content factor correlations.

`rfit_indices`, a sub-list including a variety of popular fit indices.

`rACQ_variance`, the amount of variance explained by ACQ.

`rresid_matrix`, residual matrix after partialling-out for ACQ.

`rpfactors`, factor scores for each participant.

4 Simulation studies

To assess the behavior of the proposal under favorable conditions (correct population model) and its robustness against slight misspecifications, we conducted a simulation study which focused on (a) the recovery of the 'true' loadings on both the ACQ and the content factors, and (b) the goodness of fit results.

Method

A bidimensional content model with an additional ACQ factor (see equation 1) was generated under the following specifications: (a) all the factors were orthogonal (this choice was made for the sake of simplicity); (b) the content factors contained positive and negative loadings (representing the positively and negatively keyed items), and (c) the loadings on the ACQ factor were all positive. The number of items per factor was 10 and the sample size was fixed to 300, slightly higher than recommended, to find accurate factor loading estimates (Fabrigar et al., 1999).

In the content factors, the simulated loadings had an average value of .6 (in absolute value) and a standard deviation of 0.1. For the ACQ factor the mean loading value was .2. The behavior of **siren** was assessed under three general conditions: (1) type of item: ordinal (four categories; FA based on polychoric correlations), or continuous (FA based on Pearson correlations); (2) pattern of substantive loadings at three levels: (a) completely balanced, (b) 60% of positive items, and (c) 70% of positive items; and (3) ACQ pattern at two levels: (a) equal ACQ loadings, (b) low heterogeneity (standard deviation of .01), and (c) high heterogeneity (standard deviation of .1). Thus, a factorial design with 18 experimental conditions (2 x 3 x 3) was used. These conditions were chosen according to (a) the most problematic conditions for the alternative method discussed above, and (b) the degree of realism in the applied context.

For each experimental condition, 200 replicas were generated. A higher number of replicas would simply increase the estimation time without changing the results. All analyses were conducted with R (R Core Team, 2016). The quality of the estimates was assessed using the average bias, and the consistency of the model-data fit results was assessed using an analysis of variance (ANOVA) for each fit index considered in the study (CFI, GFI, RMSR and RMSEA).

Results

The results for the average bias were very stable in all conditions (see Table 4 and 5). The loadings on content factors are recovered more accurately than the loadings on the ACQ factor. In the content factors, the bias is evenly distributed across both factors, and never exceeds .05 (continuous condition) or .04 (ordinal condition) in absolute value. In contrast, the average bias in the ACQ factor is greater in the ordinal case.

No significant changes are observed when imbalance increases. However, a slight increase in bias can be noticed in those conditions in which the ACQ pattern is more heterogeneous. This increase, however, does not substantially impact the average bias of the ACQ factor 5. The bias in ACQ remains at around .06, with a maximum of .077.

In regards of model fit, the results of ANOVA are in Table 6. None of the ANOVAs produced statistically significant results. However, when the ACQ patterns are very heterogeneous, the fit of the model tends to worsen, which suggests that **siren** is more sensitive to the pattern of ACQ loadings than to the degree of balance of the content items within each factor.

It should be noted that the simulated data are very favorable, each factor has more than enough

Table 4: Raw bias of the content factors

		factor 1										factor 2									
Continuous		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Bal	EQ	.047	.044	.045	.047	.045	.043	.045	.042	.04	.041	.044	.043	.042	.046	.045	.046	.041	.049	.042	.044
Bal	LH	.047	.044	.045	.047	.045	.043	.045	.042	.046	.041	.047	.043	.044	.043	.045	.043	.041	.046	.04	.041
Bal	HH	.046	.043	.047	.047	.045	.044	.045	.045	.043	.044	.045	.046	.04	.046	.043	.043	.044	.05	.045	.042
LU	EQ	.04	.044	.047	.044	.043	.046	.044	.042	.043	.043	.04	.043	.045	.045	.045	.042	.043	.044	.042	.043
LU	LH	.044	.046	.043	.044	.044	.043	.046	.043	.043	.045	.045	.041	.041	.039	.043	.042	.043	.043	.043	.042
LU	HH	.043	.043	.038	.043	.043	.046	.043	.042	.046	.05	.046	.044	.045	.044	.042	.044	.042	.043	.041	.039
HU	EQ	.044	.047	.041	.041	.048	.046	.045	.042	.049	.05	.044	.046	.044	.045	.044	.046	.046	.049	.041	.046
HU	LH	.043	.045	.046	.045	.045	.041	.045	.04	.043	.045	.048	.045	.046	.043	.048	.04	.043	.045	.043	.045
HU	HH	.05	.044	.046	.045	.047	.043	.044	.047	.05	.049	.046	.049	.042	.048	.046	.047	.049	.045	.044	.046
Ordinal		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Bal	EQ	.04	.036	.036	.039	.039	.038	.036	.037	.035	.041	.038	.039	.037	.038	.035	.035	.036	.04	.037	.031
Bal	LH	.04	.036	.036	.039	.039	.038	.036	.037	.035	.041	.039	.042	.037	.036	.035	.034	.038	.039	.037	.034
Bal	HH	.036	.036	.036	.038	.034	.037	.037	.037	.036	.036	.037	.035	.036	.04	.04	.037	.037	.038	.038	.037
LU	EQ	.039	.035	.037	.04	.034	.038	.035	.035	.033	.036	.032	.038	.038	.035	.036	.036	.04	.035	.037	.037
LU	LH	.039	.034	.038	.038	.037	.037	.037	.034	.04	.039	.04	.037	.035	.033	.039	.034	.036	.037	.039	.034
LU	HH	.037	.037	.039	.04	.036	.038	.034	.037	.038	.039	.035	.037	.037	.037	.04	.04	.037	.036	.037	.033
HU	EQ	.036	.036	.031	.039	.037	.037	.036	.041	.039	.037	.035	.04	.037	.033	.036	.035	.038	.044	.04	.042
HU	LH	.037	.038	.036	.04	.037	.035	.036	.036	.041	.038	.036	.038	.035	.035	.041	.037	.036	.036	.039	.039
HU	HH	.038	.037	.033	.042	.038	.038	.039	.036	.035	.038	.04	.04	.039	.038	.038	.043	.035	.038	0.037	.038

Note: Bal = Balanced; LU = Low Unbalanced; HU = High Unbalanced, EQ = equal; LH = Low Heterogeneity, HH = High Heteregeneity

Table 5: Raw bias of the ACQ factor

		factor 1										factor 2									
Continuous		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Bal	EQ	.063	.066	.064	.074	.068	.064	.063	.066	.068	.064	.07	.068	.071	.064	.064	.075	.076	.074	.068	.068
Bal	LH	.07	.069	.072	.072	.066	.075	.063	.068	.072	.073	.068	.066	.063	.068	.068	.067	.065	.067	.068	.067
Bal	HH	.063	.057	.058	.06	.061	.061	.064	.062	.061	.06	.061	.063	.065	.065	.064	.064	.062	.063	.065	.066
LU	EQ	.058	.059	.059	.056	.058	.059	.058	.066	.064	.064	.061	.064	.058	.061	.064	.061	.068	.068	.063	.07
LU	LH	.059	.061	.06	.057	.066	.06	.061	.059	.062	.072	.066	.063	.06	.063	.065	.062	.061	.069	.066	.064
LU	HH	.067	.058	.063	.059	.056	.056	.059	.062	.061	.06	.061	.063	.056	.056	.059	.054	.063	.054	.056	.059
HU	EQ	.068	0.071	0.061	0.066	.065	.059	.063	0.062	.07	.07	.072	.066	.064	.067	.077	.076	.072	.066	.068	.066
HU	LH	.063	.058	.066	.063	.064	.061	.068	.071	.068	.075	.071	.068	.069	.063	.071	.068	.074	.064	.068	.067
HU	HH	.062	.062	.064	.068	.059	.062	.063	.067	.071	.065	.066	.071	.056	.056	.071	.065	.071	.067	.067	.059
Ordinal		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Bal	EQ	.069	.075	.072	.067	.074	.075	.079	.076	.074	.073	.071	.077	.078	.076	.069	.077	.074	.068	.074	.071
Bal	LH	.075	.076	.078	.07	.079	.08	.078	.079	.069	.083	.072	.077	.079	.072	.078	.076	.071	.076	.076	.071
Bal	HH	.059	.065	.068	.073	.072	.066	.065	.066	.07	.071	.065	.073	.067	.071	.068	.065	.069	.064	.07	.066
LU	EQ	.069	.062	.067	.07	.069	.072	.073	.073	.076	.07	.07	.066	.069	.068	.071	.072	.079	.072	.071	.072
LU	LH	.073	.061	.074	.071	.074	.073	.068	.076	.073	.065	.069	.076	.065	.073	.073	.074	.075	.072	.072	.075
LU	HH	.064	.069	.071	.065	.072	.072	.074	.068	.073	.076	.067	.068	.063	.068	.067	.07	.078	.06	.069	.064
HU	EQ	.076	.071	.067	.073	.07	.074	.073	.08	.069	.089	.081	.075	.078	.075	.073	.077	.076	.085	.074	.074
HU	LH	.072	.069	.07	.077	.078	.068	.068	.086	.079	.078	.078	.077	.074	.081	.074	.08	.077	.083	.082	.083
HU	HH	.06	.069	.073	.074	.07	.071	.076	.078	.077	.074	.077	.063	.076	.067	.067	.069	.08	.071	.076	.074

Note: Bal = Balanced; LU = Low Unbalanced; HU = High Unbalanced, EQ = equal; LH = Low Heterogeneity, HH = High Heteregeneity

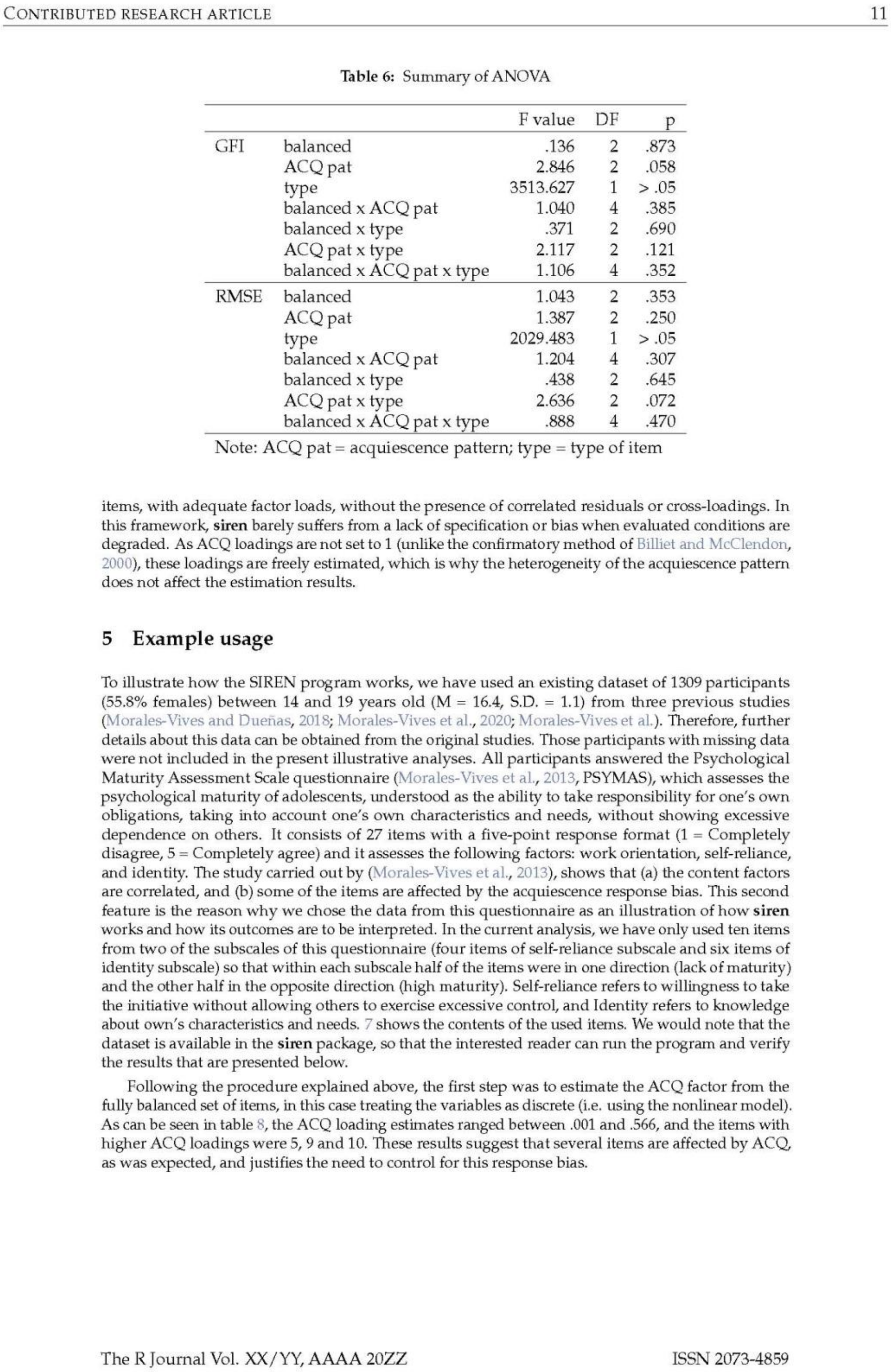


Table 6: Summary of ANOVA

		F value	DF	p
GFI	balanced	.136	2	.873
	ACQ pat	2.846	2	.058
	type	3513.627	1	> .05
	balanced x ACQ pat	1.040	4	.385
	balanced x type	.371	2	.690
	ACQ pat x type	2.117	2	.121
	balanced x ACQ pat x type	1.106	4	.352
RMSE	balanced	1.043	2	.353
	ACQ pat	1.387	2	.250
	type	2029.483	1	> .05
	balanced x ACQ pat	1.204	4	.307
	balanced x type	.438	2	.645
	ACQ pat x type	2.636	2	.072
	balanced x ACQ pat x type	.888	4	.470

Note: ACQ pat = acquiescence pattern; type = type of item

items, with adequate factor loads, without the presence of correlated residuals or cross-loadings. In this framework, **siren** barely suffers from a lack of specification or bias when evaluated conditions are degraded. As ACQ loadings are not set to 1 (unlike the confirmatory method of Billiet and McClendon, 2000), these loadings are freely estimated, which is why the heterogeneity of the acquiescence pattern does not affect the estimation results.

5 Example usage

To illustrate how the SIREN program works, we have used an existing dataset of 1309 participants (55.8% females) between 14 and 19 years old (M = 16.4, S.D. = 1.1) from three previous studies (Morales-Vives and Dueñas, 2018; Morales-Vives et al., 2020; Morales-Vives et al.). Therefore, further details about this data can be obtained from the original studies. Those participants with missing data were not included in the present illustrative analyses. All participants answered the Psychological Maturity Assessment Scale questionnaire (Morales-Vives et al., 2013, PSYMAS), which assesses the psychological maturity of adolescents, understood as the ability to take responsibility for one's own obligations, taking into account one's own characteristics and needs, without showing excessive dependence on others. It consists of 27 items with a five-point response format (1 = Completely disagree, 5 = Completely agree) and it assesses the following factors: work orientation, self-reliance, and identity. The study carried out by (Morales-Vives et al., 2013), shows that (a) the content factors are correlated, and (b) some of the items are affected by the acquiescence response bias. This second feature is the reason why we chose the data from this questionnaire as an illustration of how **siren** works and how its outcomes are to be interpreted. In the current analysis, we have only used ten items from two of the subscales of this questionnaire (four items of self-reliance subscale and six items of identity subscale) so that within each subscale half of the items were in one direction (lack of maturity) and the other half in the opposite direction (high maturity). Self-reliance refers to willingness to take the initiative without allowing others to exercise excessive control, and Identity refers to knowledge about own's characteristics and needs. 7 shows the contents of the used items. We would note that the dataset is available in the **siren** package, so that the interested reader can run the program and verify the results that are presented below.

Following the procedure explained above, the first step was to estimate the ACQ factor from the fully balanced set of items, in this case treating the variables as discrete (i.e. using the nonlinear model). As can be seen in table 8, the ACQ loading estimates ranged between .001 and .566, and the items with higher ACQ loadings were 5, 9 and 10. These results suggest that several items are affected by ACQ, as was expected, and justifies the need to control for this response bias.

Table 7: Loading estimates in the Acquiescence factor obtained in the first step

	ACQ
Item 1. Consult the peer group before buying clothes	.001
Item 2. Friends' opinions determine what is considered wrong	.001
Item 6. Doesn't mind doing different things than friends	.231
Item 9. Facing consequences of one's own mistakes	.206
Item 3. Not showing the true self	.380
Item 4. Feeling accepted and valued	.079
Item 5. Feeling empty	.070
Item 7. Good self-knowledge	.156
Item 8. Others do not really know him/her	.338
Item 10. Feeling capable of doing many things well	.566

Note: ACQ = Acquiescence

We next fitted a CFA solution consisting of two correlated content factors with a full IC structure, in which each item only had a non-zero loading on the own factor, and an additional ACQ factor in which the corresponding loading was fixed at the estimate obtained in table A (in the ordinal case there is no need to multiply this loading by the standard deviation as this has a unit value). The final ULS estimates for the full solution are in table 8. As expected, the four items of self-reliance subscale loaded in one factor, and the six items of the identity subscale loaded in the other factor. Inspection of the signs of the loadings suggests that the full condition of balance within each factor is achieved. As for the strength of the content solution, items 2, 6, 9 had loadings of .40 or higher (in absolute value) on the self-reliance factor, while item 1 had the lowest loading, the same result obtained in the study carried out by Morales-Vives et al. (2013). All the items of identity had loadings on this factor higher than .40, being item 5 the item with the highest loading, which, again, agrees with the results by Morales-Vives et al. (2013). Overall, the procedures included in the **siren** program provide the expected results, which are congruent with those reported in the previous study, even though the latter included a greater number of items than in the present study. Furthermore, the correlation between the two factors is .436, as was expected, because the study carried out by Morales-Vives et al. (2013) already showed that these factors are positively correlated.

Table 8: Loadings obtained in the CFA analysis

	Factor 1	Factor 2	ACQ
Item 1. Consult the peer group before buying clothes	.29	.00	.001
Item 2. Friends' opinions determine what is considered wrong	.53	.00	.001
Item 6. Doesn't mind doing different things than friends	-.56	.00	.231
Item 9. Facing consequences of one's own mistakes	-.40	.00	.206
Item 3. Not showing the true self	.00	.54	.380
Item 4. Feeling accepted and valued	.00	-.57	.079
Item 5. Feeling empty	.00	.66	.070
Item 7. Good self-knowledge	.00	-.53	.156
Item 8. Others do not really know him/her	.00	.43	.338
Item 10. Feeling capable of doing many things well	.00	-.47	.566

Note: ACQ = Acquiescence

The fit of the solution on 8 was quite acceptable: GFI=.99, RMSR=.04, RMSEA=0.04, and CFI=0.96. This good fit suggests that, once ACQ is controlled, the structure of the PSYMAS item pool assessed here is remarkably simple and strong.

6 Concluding remarks

There are at present two factor-analytic approaches for calibrating and scoring typical-response measures after controlling for the biasing effects of AR. One of them is fully confirmatory, and the complete solution is identified by fixing all the ACQ loadings to the same value. The other is unrestricted (i.e. exploratory or semi-confirmatory). According to the literature, each of the two approaches has their pros and cons (Savalei and Falk, 2014; de la Fuente and Abad, 2020)

In this article we have proposed a hybrid EFA-CFA procedure, called SIREN, that tries to combine the best features of the two approaches above. Thus, in SIREN, the ACQ factor can be identified in a first step without the need to constrain all its loadings to have the same value. Next, once the ACQ factor is identified, a fully confirmatory (restricted) solution can be specified for the content factors at the second step. Finally, for both types of factors (ACQ and content), our proposed procedure allows factor score estimates for each individual to be obtained at the third step. The flexibility of what we propose widens the available options for assessing the structural properties of the typical-response measure under scrutiny, and also, for obtaining accurate score estimates for each individual. Regarding this last point, we would note that most existing factor-analytical developments designed for controlling ACQ tend to focus solely on the structural properties of the instrument. However, accurate and ”clean” individual score estimates might be highly relevant in further validity studies or if clinical decisions have to be taken on the basis of this instrument.

Apart from increased flexibility, the proposal has many features that considerably increase its range of application. To start with, it allows solutions to be fitted with the standard linear FA model or with the non-linear graded-response model. Second, the solution can be fitted using a ”cleaned” residual covariance matrix (the standard approach to this type of problems) or directly fitted to the raw data using the ACQ loading estimates as fixed and known. This second option makes it possible to use a wide range of estimation procedures and goodness of fit measures for estimating and assessing model data fit.

The main theoretical and potential shortcoming of SIREN is the loss of efficiency caused by the sequential limited-information procedure which it uses. So far, the results of the simulation study suggest that this loss has little impact in practice. However, more extensive simulation is warranted.

The R program that implements SIREN (and which has the same name) has been designed to be as user-friendly as possible, and requires very few specifications from the user: essentially, the FA model of choice (linear or nonlinear) and a target matrix, which specifies the content factor on which each item is expected to load together with the expected sign of this loading. So, the program can be used by practitioners with minimal proficiency in FA. Furthermore, **siren** is extremely versatile, and provides a considerable amount of information in an output that is simple and clear to interpret.

Bibliography

F. M. Andrews. Construct validity and error components of survey measures: A structural modeling approach. *Public opinion quarterly*, 2(48):409–442, 1984. URL https://doi.org/10.1086/268840. [p3]

J. B. Billiet and M. J. McClendon. Modeling acquiescence in measurement models for two balanced sets of items. *Structural equation modeling*, 4(7):608–628, 2000. URL https://doi.org/10.1207/S15328007SEM0704_5. [p1, 3, 11]

K. A. Bollen. A new incremental fit index for general structural equation models. *Sociological methods and research*, 3(17):303–316, 1989. URL https://doi.org/10.1177/0049124189017003004. [p4]

V. Chulakian. The optimality of the centroid method. *Psychometrika*, 3(68):473–475, 2003. URL https://doi.org/10.1007/BF02294738. [p4]

J. de la Fuente and F. J. Abad. Comparing methods for modeling acquiescence in multidimensional partially balanced scales. *Psicothema*, 4(32):590–597, 2020. URL http://10.7334/psicothema2020.96. [p1, 12]

A. DeCastellarnau and W. E. Saris. Correcting correlation and covariance matrices for measurement errors before further analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, 4(28):572–581, 2021. URL https://doi.org/10.1080/10705511.2020.1870229. [p3, 7]

C. E. DeMars. An illustration of the effects of ignoring a secondary factor. *Applied Psychological Measurement*, 38(5):406–409, 2014. URL https://doi.org/10.1177/0146621614529360. [p2, 5]

H. J. Eysenck. Criterion analysis–an application of the hypothetico-deductive method to factor analysis. *Psychological Review*, 1(57):38–53, 1950. URL https://doi.org/10.1037/h0057657. [p4]

L. R. Fabrigar, D. T. Wegener, R. C. MacCallum, and E. J. Strahan. Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 3(4):272–299, 1999. URL https://doi.org/10.1037/1082-989X.4.3.272. [p8]

P. J. Ferrando and U. Lorenzo-Seva. Acquiescence as a source of bias and model and person misfit: A theoretical and empirical analysis. *British Journal of Mathematical and Statistical Psychology*, 2(62): 427–448, 2010. URL https://doi.org/10.1348/000711009X470740. [p2, 4, 5]

P. J. Ferrando and U. Lorenzo-Seva. Unrestricted item factor analysis and some relations with item response theory. Technical report, Department of Psychology, Universitat Rovira i Virgili, Tarragona, 2013. URL `http://psico.fcep.urv.es/utilitats/factor.` [p4]

P. J. Ferrando and U. Lorenzo-Seva. Unrestricted versus restricted factor analysis of multidimensional test items: some aspects of the problem and some suggestions. *Psicologica*, 2(37):235–247, 2016. URL `https://www.redalyc.org/articulo.oa?id=16946248007`. [p5]

P. J. Ferrando, U. Lorenzo-Seva, and E. Chico. Unrestricted factor analytic procedures for assessing acquiescent responding in balanced, theoretically unidimensional personality scales. *Multivariate Behavioral Research*, 3(38):353–374, 2003. URL `https://doi.org/10.1207/S15327906MBR3803_04`. [p1, 3]

K. G. Joreskog. A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, 2(34):183–202. [p2]

D. Lawley. Approximate methods in factor analysis. *British Journal of Statistical Psychology*, 13(1):11–17, 1960. [p4]

U. Lorenzo-Seva and P. J. Ferrando. Acquiescent responding in partially balanced multidimensional scales. *British Journal of Mathematical and Statistical Psychology*, 2(62):319–326, 2009. URL `https://doi.org/10.1348/000711007X265164`. [p3]

R. P. McDonald. A simple comprehensive model for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, 1(31):59–72, 1978. URL `https://doi.org/10.1111/j.2044-8317.1978.tb00573.x`. [p4, 6]

R. P. McDonald. A basis for multidimensional item response theory. *Applied Psychological Measurement*, 2(24):99–114, 2000. URL `https://doi.org/10.1177/01466210022031552`. [p2]

R. P. McDonald and M. M. Mok. Goodness of fit in item response models. *Multivariate Behavioral Research*, 1(30):23–40, 1995. URL `https://doi.org/10.1207/s15327906mbr3001_2`. [p5]

S. Messick. The psychology of acquiescence: an interpretation of research evidence 1. *ETS Research Bulletin Series*, 1966(1):i–44, 1966. [p1]

F. Morales-Vives and J. M. Dueñas. Predicting suicidal ideation in adolescent boys and girls: The role of psychological maturity, personality traits, depression and life satisfaction. *The Spanish journal of psychology*, 21:E10, 2018. URL `https://doi.org/10.1017/sjp.2018.12`. [p11]

F. Morales-Vives, P. Ferrando, J. M. Dueñas, S. Martín-Arbós, and E. Castarlenas. Are older teens more frustrated than younger teens by the covid-19 restrictions? the role of psychological maturity, personality traits, depression and life satisfaction. *Current Psychology*. [p11]

F. Morales-Vives, E. Camps, and U. Lorenzo-Seva. Development and validation of the psychological maturity assessment scale (psymas). *European Journal of Psychological Assessment*, 1(29):12–18, 2013. URL `https://doi.org/10.1027/1015-5759/a000115`. [p11, 12]

F. Morales-Vives, E. Camps, and J. M. Dueñas. Predicting academic achievement in adolescents: The role of maturity, intelligence and personality. *Psicothema*, 1(31):84–91, 2020. URL `https://doi.org/10.1027/1015-5759/a000115`. [p11]

B. Muthen. Goodness of fit with categorical and other non-normal variables. In K. Bollen and S. J. Long, editors, *Testing Structural Equation Models*, pages 205–243. Sage Publications, 1993. [p4]

J. C. Nunnally. An overview of psychological measurement. In B. Wolman, editor, *Clinical diagnosis of mental disorders*, pages 97–146. Springer, 1978. [p1, 7]

D. L. Oberski and A. Satorra. Measurement error models with uncertainty about the error variance. *Structural Equation Modeling: A Multidisciplinary Journal*, 3(20):409–428, 2013. URL `https://doi.org/10.1080/10705511.2013.797820`. [p7]

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL `https://www.R-project.org/`. ISBN 3-900051-07-0. [p8]

V. Savalei and C. F. Falk. Recovering substantive factor loadings in the presence of acquiescence bias: A comparison of three approaches. *Multivariate behavioral research*, 5(49):407–424, 2014. URL `https://doi.org/10.1037/1082-989X.4.3.272`. [p1, 12]

J. S. Tanaka. An overview of psychological measurement. In K. Bollen and S. J. Long, editors, *Testing Structural Equation Models*, pages 10–40. Sage Publications, 1993. [p5]

J. M. F. ten Berge. A legitimate case of component analysis of ipsative measures, and partialling the mean as an alternative to ipsatization. *Multivariate Behavioral Research*, 4(34):89–102, 2020. URL https://doi.org/10.1207/s15327906mbr3401_4. [p3, 4, 6, 7]

J. M. F. ten Berge and H. A. L. Kiers. A numerical approach to the approximate and the exact minimum rank of a covariance matrix. *Psychometrika*, 2(56):309–215, 1991. URL https://doi.org/10.1007/BF02294464. [p4]

T. Asparouhov and B. Muthen. Residual structural equation models. *Structural Equation Modeling*, 1 (30):1–31, 1984. URL https://doi.org/10.1080/10705511.2022.2074422. [p3]

A. Vigil-Colet, D. Navarro-Gonzalez, and F. Morales-Vives. To reverse or to not reverse likert-type items: That is the question. *Psicothema*, 1(32):108–114, 2020. URL https://doi.org/10.7334/psicothema2019.286. [p1]

David Navarro-Gonzalez
Department of Psychology
University of Lleida
Spain
0000-0002-9843-5058
david.navarro@udl.cat

Pere J. Ferrando
Department of Psychology
University Rovira i Virgili
Spain
0000-0002-3133-5466
perejoan.ferrando@urv.cat

Fabia Morales-Vives
Department of Psychology
University Rovira i Virgili
Spain
0000-0002-2095-0244
fabia.morales@urv.cat

Ana Hernandez-Dorado
Department of Psychology
University Rovira i Virgili
Spain
0000-0001-9502-9735
ana.hernandez@urv.cat

STRUCTURAL EQUATION MODELING: A MULTIDISCIPLINARY JOURNAL
2022, VOL. 29, NO. 4, 630–638
https://doi.org/10.1080/10705511.2021.2004543

Routledge
Taylor & Francis Group

OPEN ACCESS Check for updates

Detecting Correlated Residuals in Exploratory Factor Analysis: New Proposals and a Comparison of Procedures

Pere J. Ferrando, Ana Hernandez-Dorado, and Urbano Lorenzo-Seva

Universitat Rovira I Virgili

ABSTRACT
In the classical exploratory factor analysis (EFA) model, residuals are constrained to be uncorrelated. However, since the 1960s, extensions of the classical model that allow correlated residuals to be modeled exist. Furthermore, in many EFA applications (especially those intended for item analysis) it is highly relevant to decide whether an extended solution is more appropriate than the simpler classical solution. This decision, in turn, requires effective and powerful methods for detecting correlated residuals (doublets) when they are really present to be available. This paper discusses two existing detection approaches in the EFA context, and proposes a third, new procedure. Reference values, based on the concept of parallel analysis, are proposed for deciding the relevance of the flagged doublets in all the considered procedures. The functioning of the three procedures is assessed by using simulation, and illustrated with an illustrative example. The proposal, finally, has been implemented in a well-known noncommercial EFA program, and an implementation in R is being developed.

KEYWORDS
Exploratory factor analysis; correlated residuals; doublets; expected parameter change; image theory; parallel analysis

A distinctive feature of the restricted (confirmatory) factor analysis model (CFA) is that it allows correlated residuals to be specified. In contrast, in the unrestricted (exploratory) FA (EFA) model, the residual matrix is assumed to be diagonal, and so, all the residual correlations are constrained to be zero. This distinction, however, only holds for the "classical" EFA model (e.g., Mulaik, 2010; Sörbom, 1975), as, since the 1960s, more flexible EFA solutions that allow correlated residuals to be modeled have been proposed (Butler, 1968; McDonald, 1969; Mulaik, 2010; Yates, 1987). At present, correlated residuals within an EFA solution can be modeled via ESEM (Asparouhov & Muthén, 2009; Van Kesteren & Kievit, 2020), or R packages, such as the function "esem" in the "psych" package (Revelle, 2021) or combining the packages "psych" and "lavaan" (Rosseel, 2012).

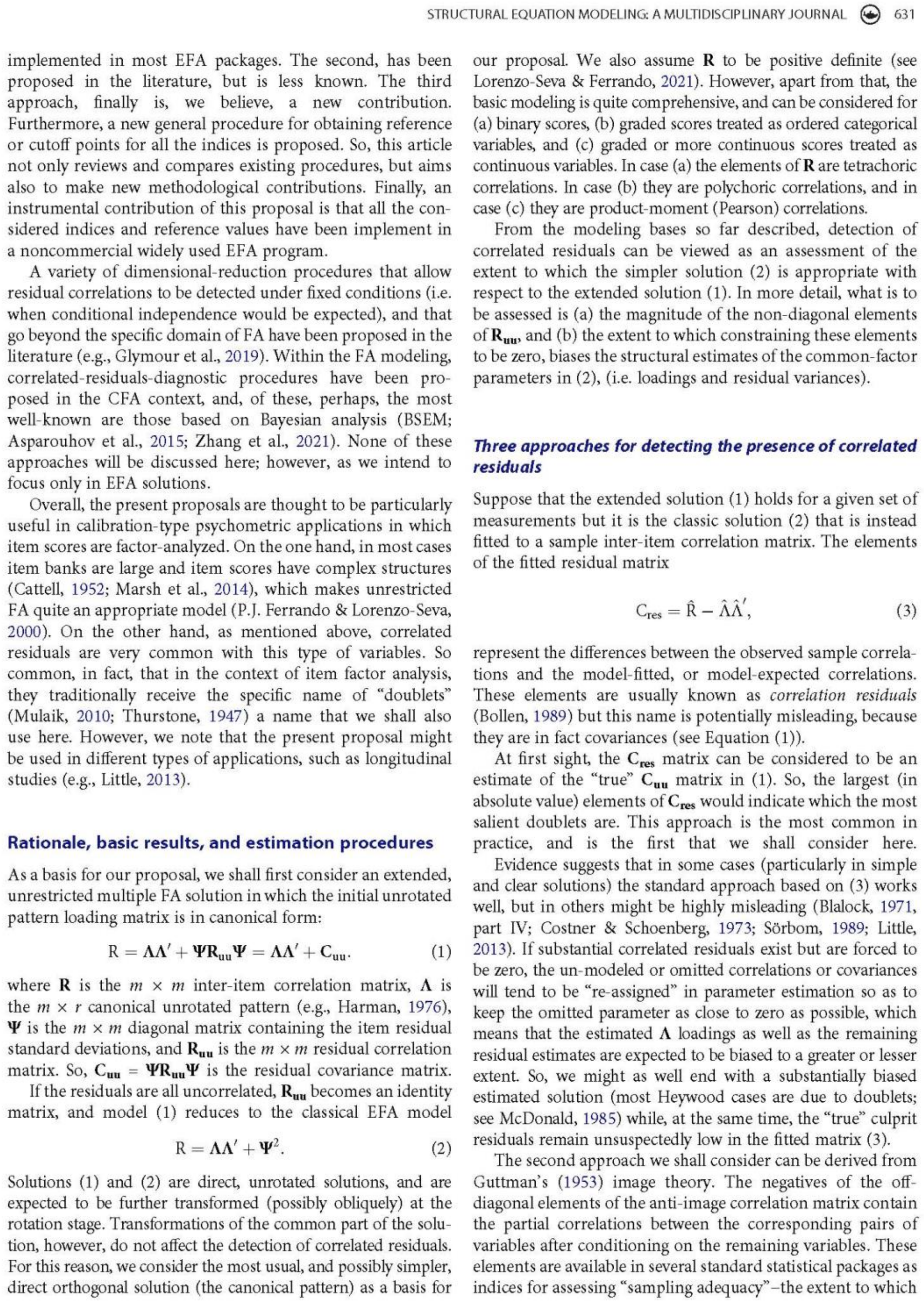
There is a vast literature and a heated debate on the convenience of allowing correlated residuals to be modeled in FA solutions (e.g., Asparouhov et al., 2015; MacCallum et al., 1992). Given the aims of this article (which are stated below), an exhaustive review of these issues is beyond its scope. However, an initial, brief discussion as well as a statement of our perspective is in order.

In an EFA based on sample data, detected correlated residuals might occur for at least two reasons: First is "true" population residuals that are perfectly justifiable (e.g., because of repeated presentation of the same items, wording similarities in the item stems, context effects … etc.). Second is sampling fluctuation around "true" zero residuals. In the first case, constraining these residuals to be zero are specification errors that, in principle, are expected to give rise to two consequences: (a) bad model-data fit, and (b) biased parameter estimates (e.g., Montoya & Edwards, 2021; Mulaik, 2010; Yates, 1987). As for point (a), however, in many applications, reasonably acceptable fits might still be obtained, but at the cost of grossly biased parameter estimates or additional factors that do not reflect substantive content (e.g., Costner & Schoenberg, 1973).

In the case of spurious correlated residuals that are due to sampling error, freeing them "blindly" with the sole purpose of improving model-data fit, is a clear case of capitalization on chance. Model-data fit in this particular sample will, no doubt, be improved, but results would not reflect "true" model-data fit in the population. Furthermore, biased estimates would be also expected to occur. However, the literature suggests that the biases in this case are not as large as those expected when "true" residuals are constrained to be zero (Reddy, 1992). We believe that blind post-hoc modification with the sole purpose of improving fit is unacceptable practice (e.g., Browne, 2001). So, the tools we shall propose in this paper are expected to be used in scenarios in which (a) samples are large enough to reasonably prevent substantial spurious correlated residuals to appear, and (b) the detected correlated residuals have a clear and defensible justification. Furthermore, we believe that cross-validation must become a routine practice for checking the consistency and stability of the diagnostic results. Having said that, we believe that obtaining efficient and powerful EFA-based diagnostic procedures for detecting correlated residuals is a relevant aim and has a clear interest.

The present article considers and compares three approaches (and indices derived from them), aimed at detecting the presence and magnitude of correlated residuals in datasets that are to be fitted by a classical EFA solution. Of these procedures, the first is the most common, and is

CONTACT Pere J. Ferrando perejoan.ferrando@urv.cat Facultad de Psicología, Universidad "Rovira I Virgili", Carretera Valls S/n, Tarragona 43007, Spain.



implemented in most EFA packages. The second, has been proposed in the literature, but is less known. The third approach, finally is, we believe, a new contribution. Furthermore, a new general procedure for obtaining reference or cutoff points for all the indices is proposed. So, this article not only reviews and compares existing procedures, but aims also to make new methodological contributions. Finally, an instrumental contribution of this proposal is that all the considered indices and reference values have been implement in a noncommercial widely used EFA program.

A variety of dimensional-reduction procedures that allow residual correlations to be detected under fixed conditions (i.e. when conditional independence would be expected), and that go beyond the specific domain of FA have been proposed in the literature (e.g., Glymour et al., 2019). Within the FA modeling, correlated-residuals-diagnostic procedures have been proposed in the CFA context, and, of these, perhaps, the most well-known are those based on Bayesian analysis (BSEM; Asparouhov et al., 2015; Zhang et al., 2021). None of these approaches will be discussed here; however, as we intend to focus only in EFA solutions.

Overall, the present proposals are thought to be particularly useful in calibration-type psychometric applications in which item scores are factor-analyzed. On the one hand, in most cases item banks are large and item scores have complex structures (Cattell, 1952; Marsh et al., 2014), which makes unrestricted FA quite an appropriate model (P.J. Ferrando & Lorenzo-Seva, 2000). On the other hand, as mentioned above, correlated residuals are very common with this type of variables. So common, in fact, that in the context of item factor analysis, they traditionally receive the specific name of "doublets" (Mulaik, 2010; Thurstone, 1947) a name that we shall also use here. However, we note that the present proposal might be used in different types of applications, such as longitudinal studies (e.g., Little, 2013).

Rationale, basic results, and estimation procedures

As a basis for our proposal, we shall first consider an extended, unrestricted multiple FA solution in which the initial unrotated pattern loading matrix is in canonical form:

$$\mathrm{R} = \boldsymbol{\Lambda\Lambda}' + \boldsymbol{\Psi}\mathbf{R}_{\mathbf{uu}}\boldsymbol{\Psi} = \boldsymbol{\Lambda\Lambda}' + \mathbf{C}_{\mathbf{uu}}. \tag{1}$$

where $\mathbf{R}$ is the $m \times m$ inter-item correlation matrix, $\boldsymbol{\Lambda}$ is the $m \times r$ canonical unrotated pattern (e.g., Harman, 1976), $\boldsymbol{\Psi}$ is the $m \times m$ diagonal matrix containing the item residual standard deviations, and $\mathbf{R}_{\mathbf{uu}}$ is the $m \times m$ residual correlation matrix. So, $\mathbf{C}_{\mathbf{uu}} = \boldsymbol{\Psi}\mathbf{R}_{\mathbf{uu}}\boldsymbol{\Psi}$ is the residual covariance matrix.

If the residuals are all uncorrelated, $\mathbf{R}_{\mathbf{uu}}$ becomes an identity matrix, and model (1) reduces to the classical EFA model

$$\mathrm{R} = \boldsymbol{\Lambda\Lambda}' + \boldsymbol{\Psi}^2. \tag{2}$$

Solutions (1) and (2) are direct, unrotated solutions, and are expected to be further transformed (possibly obliquely) at the rotation stage. Transformations of the common part of the solution, however, do not affect the detection of correlated residuals. For this reason, we consider the most usual, and possibly simpler, direct orthogonal solution (the canonical pattern) as a basis for our proposal. We also assume $\mathbf{R}$ to be positive definite (see Lorenzo-Seva & Ferrando, 2021). However, apart from that, the basic modeling is quite comprehensive, and can be considered for (a) binary scores, (b) graded scores treated as ordered categorical variables, and (c) graded or more continuous scores treated as continuous variables. In case (a) the elements of $\mathbf{R}$ are tetrachoric correlations. In case (b) they are polychoric correlations, and in case (c) they are product-moment (Pearson) correlations.

From the modeling bases so far described, detection of correlated residuals can be viewed as an assessment of the extent to which the simpler solution (2) is appropriate with respect to the extended solution (1). In more detail, what is to be assessed is (a) the magnitude of the non-diagonal elements of $\mathbf{R}_{\mathbf{uu}}$, and (b) the extent to which constraining these elements to be zero, biases the structural estimates of the common-factor parameters in (2), (i.e. loadings and residual variances).

Three approaches for detecting the presence of correlated residuals

Suppose that the extended solution (1) holds for a given set of measurements but it is the classic solution (2) that is instead fitted to a sample inter-item correlation matrix. The elements of the fitted residual matrix

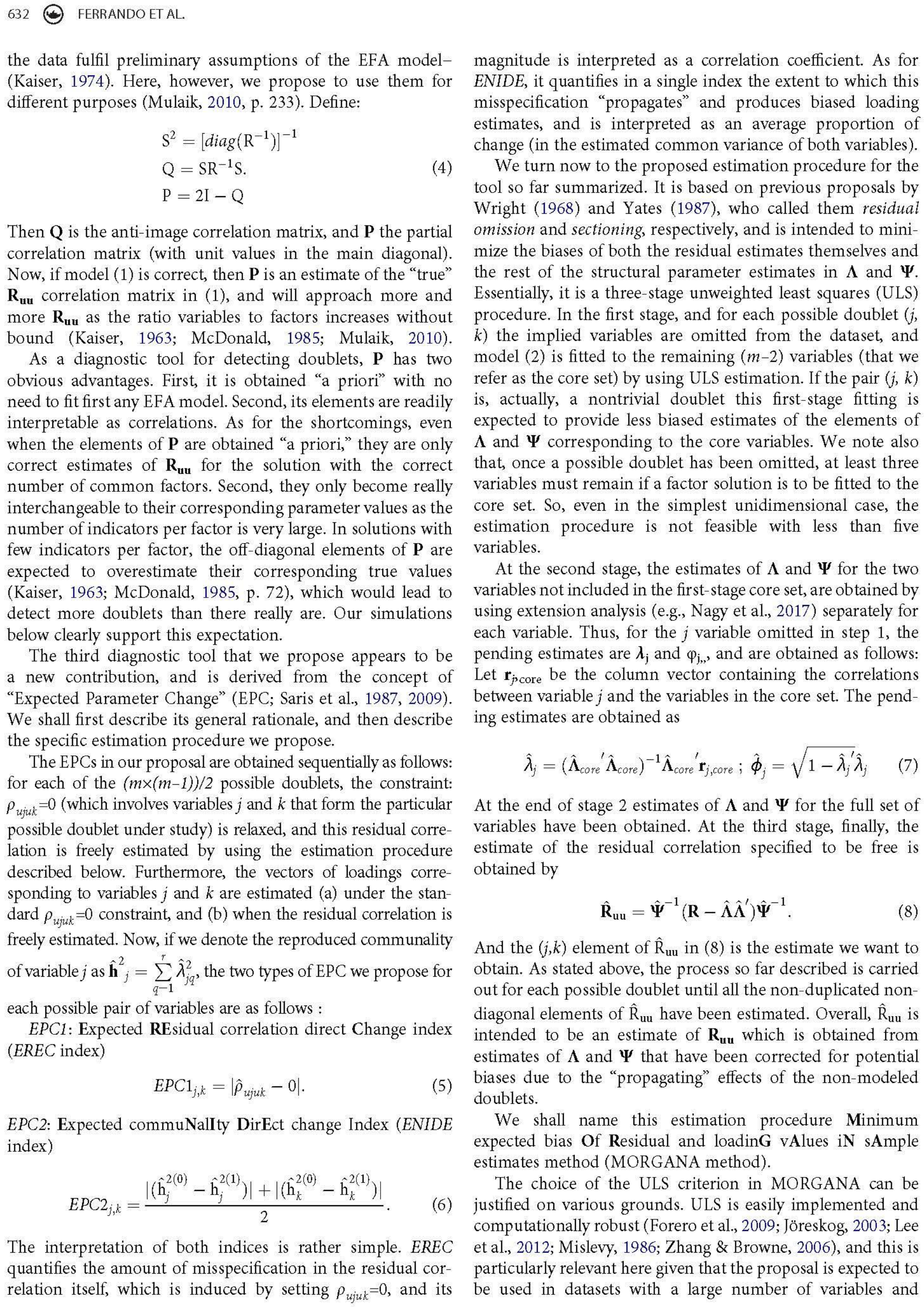
$$\mathrm{C}_{\mathbf{res}} = \hat{\mathrm{R}} - \hat{\Lambda}\hat{\Lambda}', \tag{3}$$

represent the differences between the observed sample correlations and the model-fitted, or model-expected correlations. These elements are usually known as *correlation residuals* (Bollen, 1989) but this name is potentially misleading, because they are in fact covariances (see Equation (1)).

At first sight, the $\mathbf{C}_{\mathbf{res}}$ matrix can be considered to be an estimate of the "true" $\mathbf{C}_{\mathbf{uu}}$ matrix in (1). So, the largest (in absolute value) elements of $\mathbf{C}_{\mathbf{res}}$ would indicate which the most salient doublets are. This approach is the most common in practice, and is the first that we shall consider here.

Evidence suggests that in some cases (particularly in simple and clear solutions) the standard approach based on (3) works well, but in others might be highly misleading (Blalock, 1971, part IV; Costner & Schoenberg, 1973; Sörbom, 1989; Little, 2013). If substantial correlated residuals exist but are forced to be zero, the un-modeled or omitted correlations or covariances will tend to be "re-assigned" in parameter estimation so as to keep the omitted parameter as close to zero as possible, which means that the estimated $\boldsymbol{\Lambda}$ loadings as well as the remaining residual estimates are expected to be biased to a greater or lesser extent. So, we might as well end with a substantially biased estimated solution (most Heywood cases are due to doublets; see McDonald, 1985) while, at the same time, the "true" culprit residuals remain unsuspectedly low in the fitted matrix (3).

The second approach we shall consider can be derived from Guttman's (1953) image theory. The negatives of the off-diagonal elements of the anti-image correlation matrix contain the partial correlations between the corresponding pairs of variables after conditioning on the remaining variables. These elements are available in several standard statistical packages as indices for assessing "sampling adequacy"–the extent to which



the data fulfil preliminary assumptions of the EFA model– (Kaiser, 1974). Here, however, we propose to use them for different purposes (Mulaik, 2010, p. 233). Define:

$$\begin{aligned} S^2 &= [diag(R^{-1})]^{-1} \\ Q &= SR^{-1}S. \qquad (4) \\ P &= 2I - Q \end{aligned}$$

Then **Q** is the anti-image correlation matrix, and **P** the partial correlation matrix (with unit values in the main diagonal). Now, if model (1) is correct, then **P** is an estimate of the "true" $\mathbf{R}_{\mathbf{uu}}$ correlation matrix in (1), and will approach more and more $\mathbf{R}_{\mathbf{uu}}$ as the ratio variables to factors increases without bound (Kaiser, 1963; McDonald, 1985; Mulaik, 2010).

As a diagnostic tool for detecting doublets, **P** has two obvious advantages. First, it is obtained "a priori" with no need to fit first any EFA model. Second, its elements are readily interpretable as correlations. As for the shortcomings, even when the elements of **P** are obtained "a priori," they are only correct estimates of $\mathbf{R}_{\mathbf{uu}}$ for the solution with the correct number of common factors. Second, they only become really interchangeable to their corresponding parameter values as the number of indicators per factor is very large. In solutions with few indicators per factor, the off-diagonal elements of **P** are expected to overestimate their corresponding true values (Kaiser, 1963; McDonald, 1985, p. 72), which would lead to detect more doublets than there really are. Our simulations below clearly support this expectation.

The third diagnostic tool that we propose appears to be a new contribution, and is derived from the concept of "Expected Parameter Change" (EPC; Saris et al., 1987, 2009). We shall first describe its general rationale, and then describe the specific estimation procedure we propose.

The EPCs in our proposal are obtained sequentially as follows: for each of the *(m×(m−1))/2* possible doublets, the constraint: ρ_{ujuk}=0 (which involves variables *j* and *k* that form the particular possible doublet under study) is relaxed, and this residual correlation is freely estimated by using the estimation procedure described below. Furthermore, the vectors of loadings corresponding to variables *j* and *k* are estimated (a) under the standard ρ_{ujuk}=0 constraint, and (b) when the residual correlation is freely estimated. Now, if we denote the reproduced communality of variable *j* as $\hat{\mathbf{h}}^2{}_j = \sum_{q-1}^{r} \hat{\lambda}^2_{jq}$, the two types of EPC we propose for each possible pair of variables are as follows :

EPC1: Expected **RE**sidual correlation direct **C**hange index (*EREC* index)

$$EPC1_{j,k} = |\hat{\rho}_{ujuk} - 0|. \qquad (5)$$

EPC2: Expected commu**N**al**I**ty **D**ir**E**ct change Index (*ENIDE* index)

$$EPC2_{j,k} = \frac{|(\hat{h}_j^{2(0)} - \hat{h}_j^{2(1)})| + |(\hat{h}_k^{2(0)} - \hat{h}_k^{2(1)})|}{2}. \qquad (6)$$

The interpretation of both indices is rather simple. *EREC* quantifies the amount of misspecification in the residual correlation itself, which is induced by setting ρ_{ujuk}=0, and its magnitude is interpreted as a correlation coefficient. As for *ENIDE*, it quantifies in a single index the extent to which this misspecification "propagates" and produces biased loading estimates, and is interpreted as an average proportion of change (in the estimated common variance of both variables).

We turn now to the proposed estimation procedure for the tool so far summarized. It is based on previous proposals by Wright (1968) and Yates (1987), who called them *residual omission* and *sectioning*, respectively, and is intended to minimize the biases of both the residual estimates themselves and the rest of the structural parameter estimates in $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$. Essentially, it is a three-stage unweighted least squares (ULS) procedure. In the first stage, and for each possible doublet (*j, k*) the implied variables are omitted from the dataset, and model (2) is fitted to the remaining (*m*–2) variables (that we refer as the core set) by using ULS estimation. If the pair (*j, k*) is, actually, a nontrivial doublet this first-stage fitting is expected to provide less biased estimates of the elements of $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ corresponding to the core variables. We note also that, once a possible doublet has been omitted, at least three variables must remain if a factor solution is to be fitted to the core set. So, even in the simplest unidimensional case, the estimation procedure is not feasible with less than five variables.

At the second stage, the estimates of $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ for the two variables not included in the first-stage core set, are obtained by using extension analysis (e.g., Nagy et al., 2017) separately for each variable. Thus, for the *j* variable omitted in step 1, the pending estimates are $\boldsymbol{\lambda}_j$ and $\varphi_{j,,}$ and are obtained as follows: Let $\mathbf{r}_{j,\mathrm{core}}$ be the column vector containing the correlations between variable *j* and the variables in the core set. The pending estimates are obtained as

$$\hat{\lambda}_j = (\hat{\boldsymbol{\Lambda}}_{core}{}'\hat{\boldsymbol{\Lambda}}_{core})^{-1}\hat{\boldsymbol{\Lambda}}_{core}{}'\mathbf{r}_{j,core}\ ;\ \hat{\phi}_j = \sqrt{1-\hat{\lambda}_j{}'\hat{\lambda}_j} \qquad (7)$$

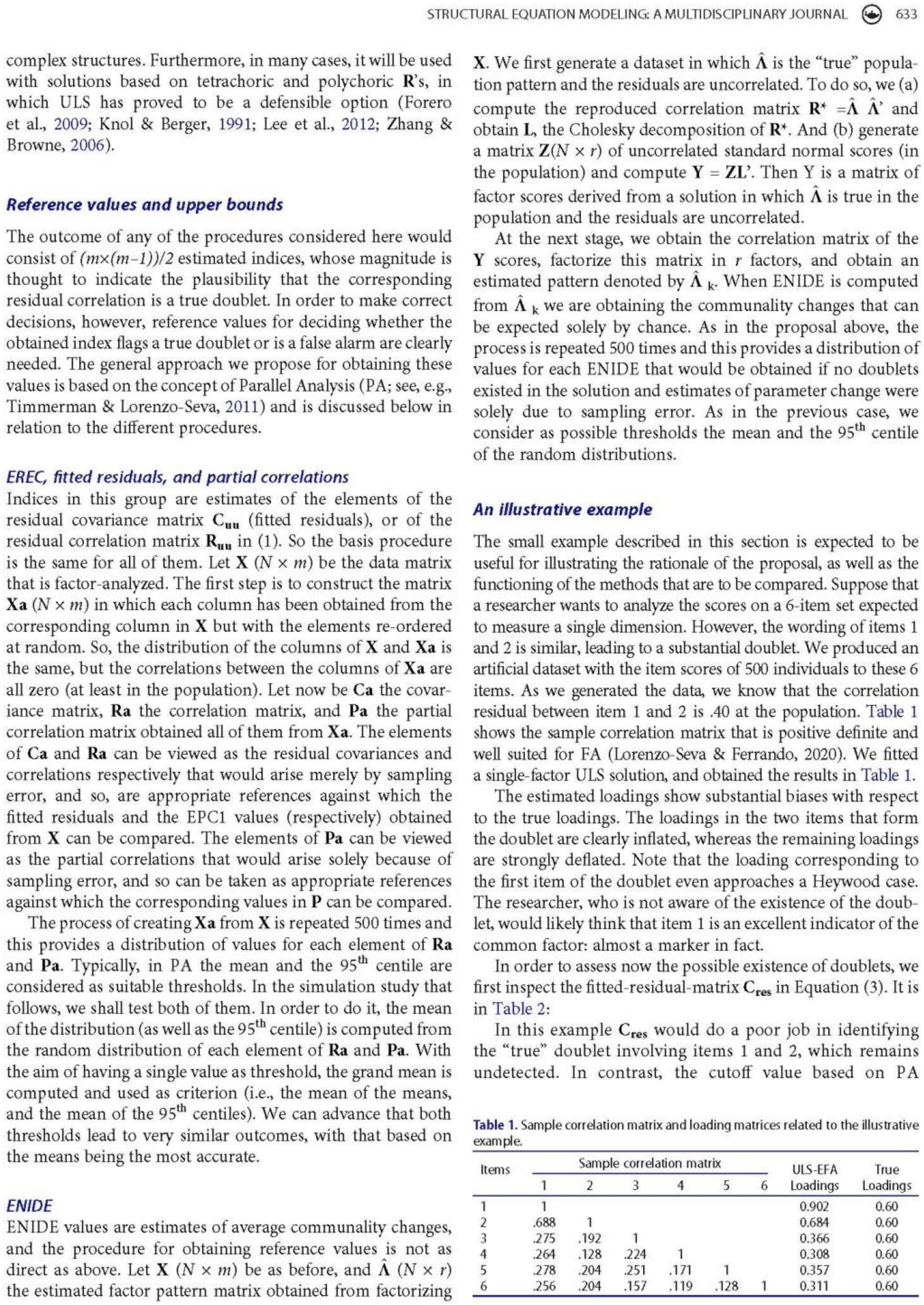
At the end of stage 2 estimates of $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ for the full set of variables have been obtained. At the third stage, finally, the estimate of the residual correlation specified to be free is obtained by

$$\hat{\mathbf{R}}_{\mathbf{uu}} = \hat{\boldsymbol{\Psi}}^{-1}(\mathbf{R} - \hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}')\hat{\boldsymbol{\Psi}}^{-1}. \qquad (8)$$

And the (*j,k*) element of $\hat{\mathrm{R}}_{\mathbf{uu}}$ in (8) is the estimate we want to obtain. As stated above, the process so far described is carried out for each possible doublet until all the non-duplicated non-diagonal elements of $\hat{\mathrm{R}}_{\mathbf{uu}}$ have been estimated. Overall, $\hat{\mathrm{R}}_{\mathbf{uu}}$ is intended to be an estimate of $\mathbf{R}_{\mathbf{uu}}$ which is obtained from estimates of $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ that have been corrected for potential biases due to the "propagating" effects of the non-modeled doublets.

We shall name this estimation procedure **M**inimum expected bias **O**f **R**esidual and loadin**G** v**A**lues i**N** s**A**mple estimates method (MORGANA method).

The choice of the ULS criterion in MORGANA can be justified on various grounds. ULS is easily implemented and computationally robust (Forero et al., 2009; Jöreskog, 2003; Lee et al., 2012; Mislevy, 1986; Zhang & Browne, 2006), and this is particularly relevant here given that the proposal is expected to be used in datasets with a large number of variables and



complex structures. Furthermore, in many cases, it will be used with solutions based on tetrachoric and polychoric **R**'s, in which ULS has proved to be a defensible option (Forero et al., 2009; Knol & Berger, 1991; Lee et al., 2012; Zhang & Browne, 2006).

Reference values and upper bounds

The outcome of any of the procedures considered here would consist of *(m×(m–1))/2* estimated indices, whose magnitude is thought to indicate the plausibility that the corresponding residual correlation is a true doublet. In order to make correct decisions, however, reference values for deciding whether the obtained index flags a true doublet or is a false alarm are clearly needed. The general approach we propose for obtaining these values is based on the concept of Parallel Analysis (PA; see, e.g., Timmerman & Lorenzo-Seva, 2011) and is discussed below in relation to the different procedures.

EREC, fitted residuals, and partial correlations

Indices in this group are estimates of the elements of the residual covariance matrix $\mathbf{C}_{\mathbf{uu}}$ (fitted residuals), or of the residual correlation matrix $\mathbf{R}_{\mathbf{uu}}$ in (1). So the basis procedure is the same for all of them. Let **X** ($N \times m$) be the data matrix that is factor-analyzed. The first step is to construct the matrix **Xa** ($N \times m$) in which each column has been obtained from the corresponding column in **X** but with the elements re-ordered at random. So, the distribution of the columns of **X** and **Xa** is the same, but the correlations between the columns of **Xa** are all zero (at least in the population). Let now be **Ca** the covariance matrix, **Ra** the correlation matrix, and **Pa** the partial correlation matrix obtained all of them from **Xa**. The elements of **Ca** and **Ra** can be viewed as the residual covariances and correlations respectively that would arise merely by sampling error, and so, are appropriate references against which the fitted residuals and the EPC1 values (respectively) obtained from **X** can be compared. The elements of **Pa** can be viewed as the partial correlations that would arise solely because of sampling error, and so can be taken as appropriate references against which the corresponding values in **P** can be compared.

The process of creating **Xa** from **X** is repeated 500 times and this provides a distribution of values for each element of **Ra** and **Pa**. Typically, in PA the mean and the 95th centile are considered as suitable thresholds. In the simulation study that follows, we shall test both of them. In order to do it, the mean of the distribution (as well as the 95th centile) is computed from the random distribution of each element of **Ra** and **Pa**. With the aim of having a single value as threshold, the grand mean is computed and used as criterion (i.e., the mean of the means, and the mean of the 95th centiles). We can advance that both thresholds lead to very similar outcomes, with that based on the means being the most accurate.

ENIDE

ENIDE values are estimates of average communality changes, and the procedure for obtaining reference values is not as direct as above. Let **X** ($N \times m$) be as before, and $\hat{\mathbf{\Lambda}}$ ($N \times r$) the estimated factor pattern matrix obtained from factorizing **X**. We first generate a dataset in which $\hat{\mathbf{\Lambda}}$ is the "true" population pattern and the residuals are uncorrelated. To do so, we (a) compute the reproduced correlation matrix $\mathbf{R}^{+} = \hat{\mathbf{\Lambda}}\ \hat{\mathbf{\Lambda}}'$ and obtain **L**, the Cholesky decomposition of $\mathbf{R}^{+}$. And (b) generate a matrix **Z**($N \times r$) of uncorrelated standard normal scores (in the population) and compute $\mathbf{Y} = \mathbf{ZL}'$. Then Y is a matrix of factor scores derived from a solution in which $\hat{\mathbf{\Lambda}}$ is true in the population and the residuals are uncorrelated.

At the next stage, we obtain the correlation matrix of the **Y** scores, factorize this matrix in *r* factors, and obtain an estimated pattern denoted by $\hat{\mathbf{\Lambda}}_k$. When ENIDE is computed from $\hat{\mathbf{\Lambda}}_k$ we are obtaining the communality changes that can be expected solely by chance. As in the proposal above, the process is repeated 500 times and this provides a distribution of values for each ENIDE that would be obtained if no doublets existed in the solution and estimates of parameter change were solely due to sampling error. As in the previous case, we consider as possible thresholds the mean and the 95th centile of the random distributions.

An illustrative example

The small example described in this section is expected to be useful for illustrating the rationale of the proposal, as well as the functioning of the methods that are to be compared. Suppose that a researcher wants to analyze the scores on a 6-item set expected to measure a single dimension. However, the wording of items 1 and 2 is similar, leading to a substantial doublet. We produced an artificial dataset with the item scores of 500 individuals to these 6 items. As we generated the data, we know that the correlation residual between item 1 and 2 is .40 at the population. Table 1 shows the sample correlation matrix that is positive definite and well suited for FA (Lorenzo-Seva & Ferrando, 2020). We fitted a single-factor ULS solution, and obtained the results in Table 1.

The estimated loadings show substantial biases with respect to the true loadings. The loadings in the two items that form the doublet are clearly inflated, whereas the remaining loadings are strongly deflated. Note that the loading corresponding to the first item of the doublet even approaches a Heywood case. The researcher, who is not aware of the existence of the doublet, would likely think that item 1 is an excellent indicator of the common factor: almost a marker in fact.

In order to assess now the possible existence of doublets, we first inspect the fitted-residual-matrix $\mathbf{C}_{\mathbf{res}}$ in Equation (3). It is in Table 2:

In this example $\mathbf{C}_{\mathbf{res}}$ would do a poor job in identifying the "true" doublet involving items 1 and 2, which remains undetected. In contrast, the cutoff value based on PA

Table 1. Sample correlation matrix and loading matrices related to the illustrative example.

Items	Sample correlation matrix						ULS-EFA	True
	1	2	3	4	5	6	Loadings	Loadings
1	1						0.902	0.60
2	.688	1					0.684	0.60
3	.275	.192	1				0.366	0.60
4	.264	.128	.224	1			0.308	0.60
5	.278	.204	.251	.171	1		0.357	0.60
6	.256	.204	.157	.119	.128	1	0.311	0.60

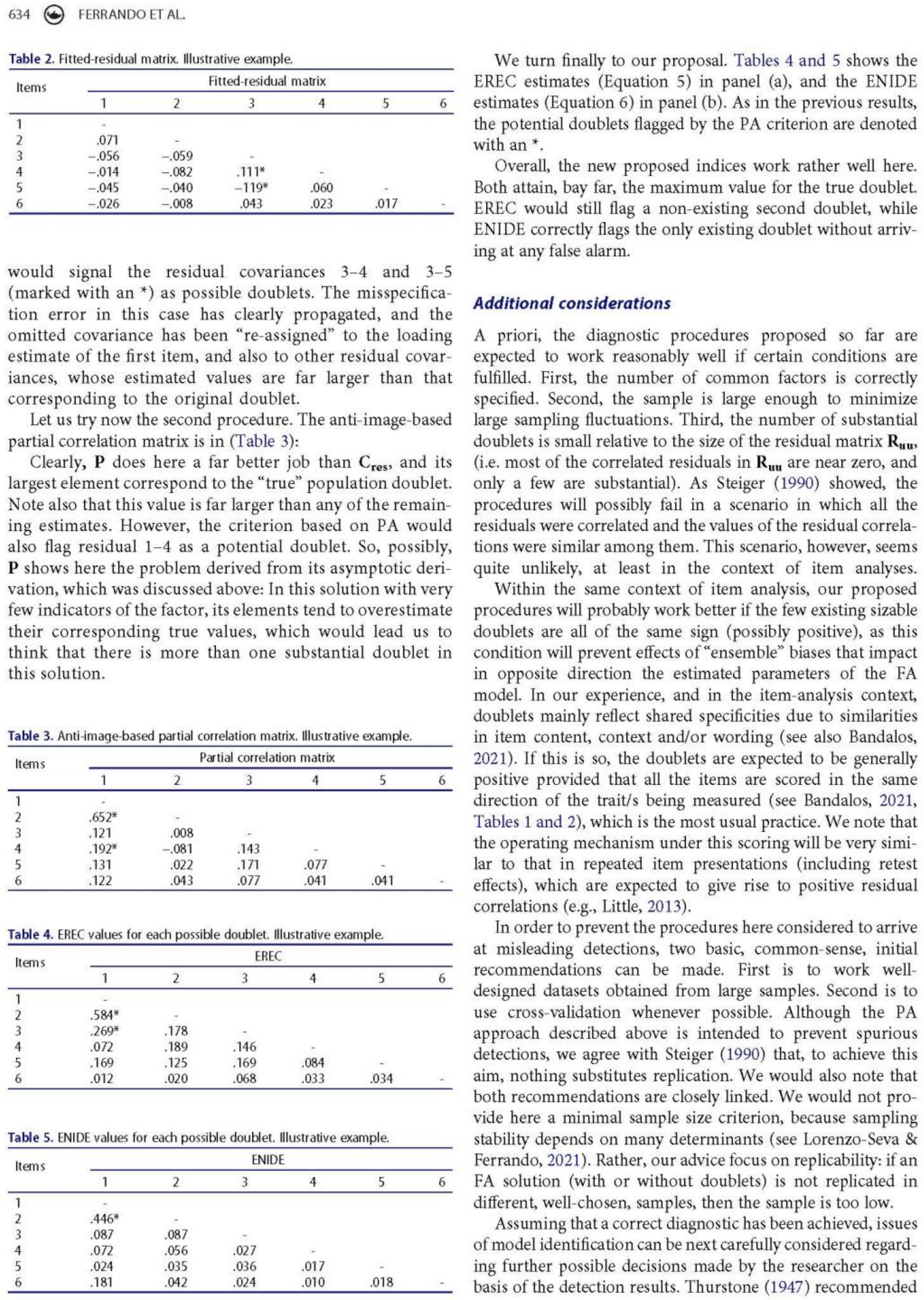


Table 2. Fitted-residual matrix. Illustrative example.

Items	Fitted-residual matrix					
	1	2	3	4	5	6
1	-					
2	.071	-				
3	−.056	−.059	-			
4	−.014	−.082	.111*	-		
5	−.045	−.040	−119*	.060	-	
6	−.026	−.008	.043	.023	.017	-

would signal the residual covariances 3–4 and 3–5 (marked with an *) as possible doublets. The misspecification error in this case has clearly propagated, and the omitted covariance has been "re-assigned" to the loading estimate of the first item, and also to other residual covariances, whose estimated values are far larger than that corresponding to the original doublet.

Let us try now the second procedure. The anti-image-based partial correlation matrix is in (Table 3):

Clearly, **P** does here a far better job than $\mathbf{C}_{\mathbf{res}}$, and its largest element correspond to the "true" population doublet. Note also that this value is far larger than any of the remaining estimates. However, the criterion based on PA would also flag residual 1–4 as a potential doublet. So, possibly, **P** shows here the problem derived from its asymptotic derivation, which was discussed above: In this solution with very few indicators of the factor, its elements tend to overestimate their corresponding true values, which would lead us to think that there is more than one substantial doublet in this solution.

Table 3. Anti-image-based partial correlation matrix. Illustrative example.

Items	Partial correlation matrix					
	1	2	3	4	5	6
1	-					
2	.652*	-				
3	.121	.008	-			
4	.192*	−.081	.143	-		
5	.131	.022	.171	.077	-	
6	.122	.043	.077	.041	.041	-

Table 4. EREC values for each possible doublet. Illustrative example.

Items	EREC					
	1	2	3	4	5	6
1	-					
2	.584*	-				
3	.269*	.178	-			
4	.072	.189	.146	-		
5	.169	.125	.169	.084	-	
6	.012	.020	.068	.033	.034	-

Table 5. ENIDE values for each possible doublet. Illustrative example.

Items	ENIDE					
	1	2	3	4	5	6
1	-					
2	.446*	-				
3	.087	.087	-			
4	.072	.056	.027	-		
5	.024	.035	.036	.017	-	
6	.181	.042	.024	.010	.018	-

We turn finally to our proposal. Tables 4 and 5 shows the EREC estimates (Equation 5) in panel (a), and the ENIDE estimates (Equation 6) in panel (b). As in the previous results, the potential doublets flagged by the PA criterion are denoted with an *.

Overall, the new proposed indices work rather well here. Both attain, bay far, the maximum value for the true doublet. EREC would still flag a non-existing second doublet, while ENIDE correctly flags the only existing doublet without arriving at any false alarm.

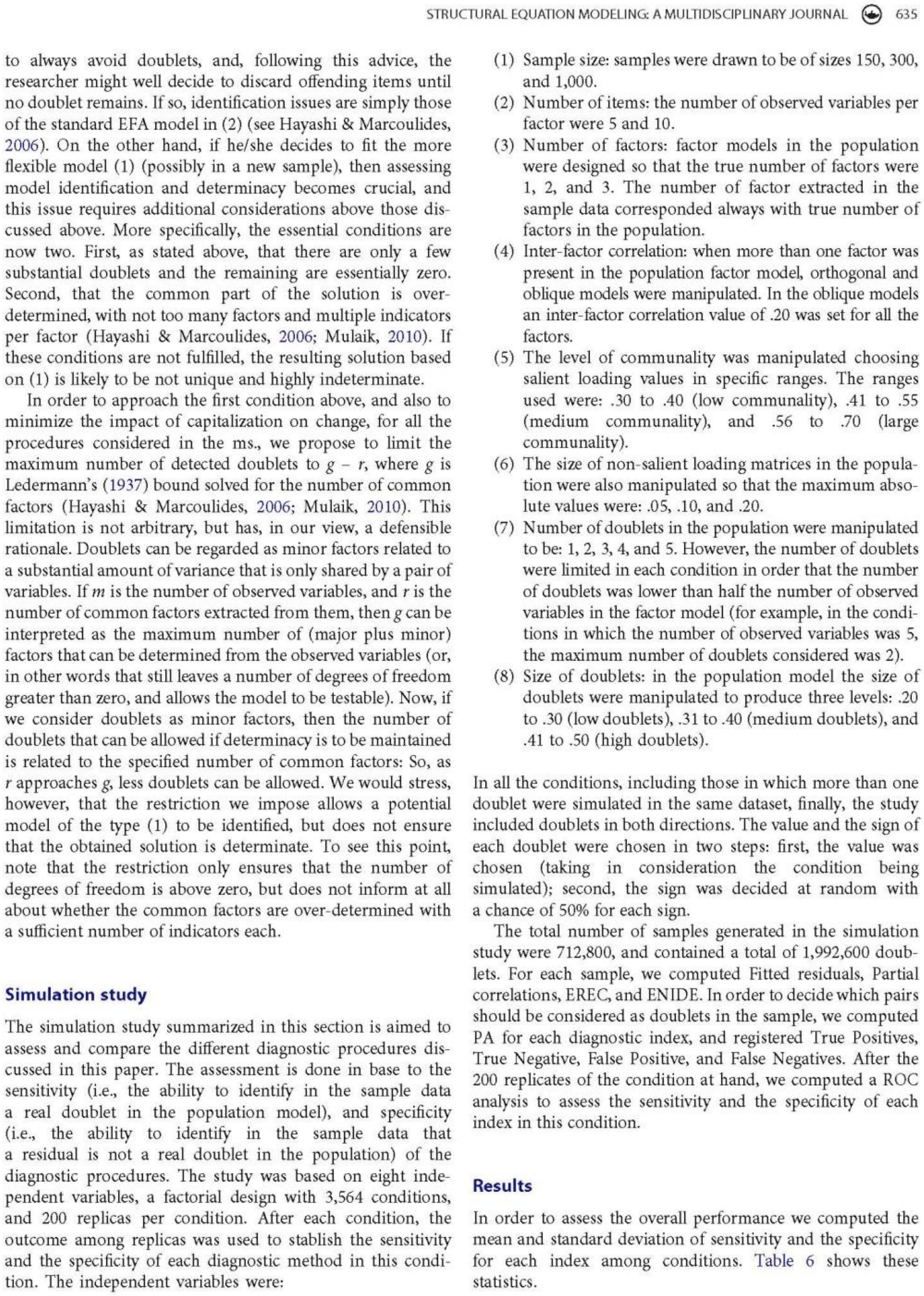
Additional considerations

A priori, the diagnostic procedures proposed so far are expected to work reasonably well if certain conditions are fulfilled. First, the number of common factors is correctly specified. Second, the sample is large enough to minimize large sampling fluctuations. Third, the number of substantial doublets is small relative to the size of the residual matrix $\mathbf{R}_{\mathbf{uu}}$, (i.e. most of the correlated residuals in $\mathbf{R}_{\mathbf{uu}}$ are near zero, and only a few are substantial). As Steiger (1990) showed, the procedures will possibly fail in a scenario in which all the residuals were correlated and the values of the residual correlations were similar among them. This scenario, however, seems quite unlikely, at least in the context of item analyses.

Within the same context of item analysis, our proposed procedures will probably work better if the few existing sizable doublets are all of the same sign (possibly positive), as this condition will prevent effects of "ensemble" biases that impact in opposite direction the estimated parameters of the FA model. In our experience, and in the item-analysis context, doublets mainly reflect shared specificities due to similarities in item content, context and/or wording (see also Bandalos, 2021). If this is so, the doublets are expected to be generally positive provided that all the items are scored in the same direction of the trait/s being measured (see Bandalos, 2021, Tables 1 and 2), which is the most usual practice. We note that the operating mechanism under this scoring will be very similar to that in repeated item presentations (including retest effects), which are expected to give rise to positive residual correlations (e.g., Little, 2013).

In order to prevent the procedures here considered to arrive at misleading detections, two basic, common-sense, initial recommendations can be made. First is to work well-designed datasets obtained from large samples. Second is to use cross-validation whenever possible. Although the PA approach described above is intended to prevent spurious detections, we agree with Steiger (1990) that, to achieve this aim, nothing substitutes replication. We would also note that both recommendations are closely linked. We would not provide here a minimal sample size criterion, because sampling stability depends on many determinants (see Lorenzo-Seva & Ferrando, 2021). Rather, our advice focus on replicability: if an FA solution (with or without doublets) is not replicated in different, well-chosen, samples, then the sample is too low.

Assuming that a correct diagnostic has been achieved, issues of model identification can be next carefully considered regarding further possible decisions made by the researcher on the basis of the detection results. Thurstone (1947) recommended



to always avoid doublets, and, following this advice, the researcher might well decide to discard offending items until no doublet remains. If so, identification issues are simply those of the standard EFA model in (2) (see Hayashi & Marcoulides, 2006). On the other hand, if he/she decides to fit the more flexible model (1) (possibly in a new sample), then assessing model identification and determinacy becomes crucial, and this issue requires additional considerations above those discussed above. More specifically, the essential conditions are now two. First, as stated above, that there are only a few substantial doublets and the remaining are essentially zero. Second, that the common part of the solution is over-determined, with not too many factors and multiple indicators per factor (Hayashi & Marcoulides, 2006; Mulaik, 2010). If these conditions are not fulfilled, the resulting solution based on (1) is likely to be not unique and highly indeterminate.

In order to approach the first condition above, and also to minimize the impact of capitalization on change, for all the procedures considered in the ms., we propose to limit the maximum number of detected doublets to $g - r$, where g is Ledermann's (1937) bound solved for the number of common factors (Hayashi & Marcoulides, 2006; Mulaik, 2010). This limitation is not arbitrary, but has, in our view, a defensible rationale. Doublets can be regarded as minor factors related to a substantial amount of variance that is only shared by a pair of variables. If m is the number of observed variables, and r is the number of common factors extracted from them, then g can be interpreted as the maximum number of (major plus minor) factors that can be determined from the observed variables (or, in other words that still leaves a number of degrees of freedom greater than zero, and allows the model to be testable). Now, if we consider doublets as minor factors, then the number of doublets that can be allowed if determinacy is to be maintained is related to the specified number of common factors: So, as r approaches g, less doublets can be allowed. We would stress, however, that the restriction we impose allows a potential model of the type (1) to be identified, but does not ensure that the obtained solution is determinate. To see this point, note that the restriction only ensures that the number of degrees of freedom is above zero, but does not inform at all about whether the common factors are over-determined with a sufficient number of indicators each.

Simulation study

The simulation study summarized in this section is aimed to assess and compare the different diagnostic procedures discussed in this paper. The assessment is done in base to the sensitivity (i.e., the ability to identify in the sample data a real doublet in the population model), and specificity (i.e., the ability to identify in the sample data that a residual is not a real doublet in the population) of the diagnostic procedures. The study was based on eight independent variables, a factorial design with 3,564 conditions, and 200 replicas per condition. After each condition, the outcome among replicas was used to stablish the sensitivity and the specificity of each diagnostic method in this condition. The independent variables were:

(1) Sample size: samples were drawn to be of sizes 150, 300, and 1,000.
(2) Number of items: the number of observed variables per factor were 5 and 10.
(3) Number of factors: factor models in the population were designed so that the true number of factors were 1, 2, and 3. The number of factor extracted in the sample data corresponded always with true number of factors in the population.
(4) Inter-factor correlation: when more than one factor was present in the population factor model, orthogonal and oblique models were manipulated. In the oblique models an inter-factor correlation value of .20 was set for all the factors.
(5) The level of communality was manipulated choosing salient loading values in specific ranges. The ranges used were: .30 to .40 (low communality), .41 to .55 (medium communality), and .56 to .70 (large communality).
(6) The size of non-salient loading matrices in the population were also manipulated so that the maximum absolute values were: .05, .10, and .20.
(7) Number of doublets in the population were manipulated to be: 1, 2, 3, 4, and 5. However, the number of doublets were limited in each condition in order that the number of doublets was lower than half the number of observed variables in the factor model (for example, in the conditions in which the number of observed variables was 5, the maximum number of doublets considered was 2).
(8) Size of doublets: in the population model the size of doublets were manipulated to produce three levels: .20 to .30 (low doublets), .31 to .40 (medium doublets), and .41 to .50 (high doublets).

In all the conditions, including those in which more than one doublet were simulated in the same dataset, finally, the study included doublets in both directions. The value and the sign of each doublet were chosen in two steps: first, the value was chosen (taking in consideration the condition being simulated); second, the sign was decided at random with a chance of 50% for each sign.

The total number of samples generated in the simulation study were 712,800, and contained a total of 1,992,600 doublets. For each sample, we computed Fitted residuals, Partial correlations, EREC, and ENIDE. In order to decide which pairs should be considered as doublets in the sample, we computed PA for each diagnostic index, and registered True Positives, True Negative, False Positive, and False Negatives. After the 200 replicates of the condition at hand, we computed a ROC analysis to assess the sensitivity and the specificity of each index in this condition.

Results

In order to assess the overall performance we computed the mean and standard deviation of sensitivity and the specificity for each index among conditions. Table 6 shows these statistics.

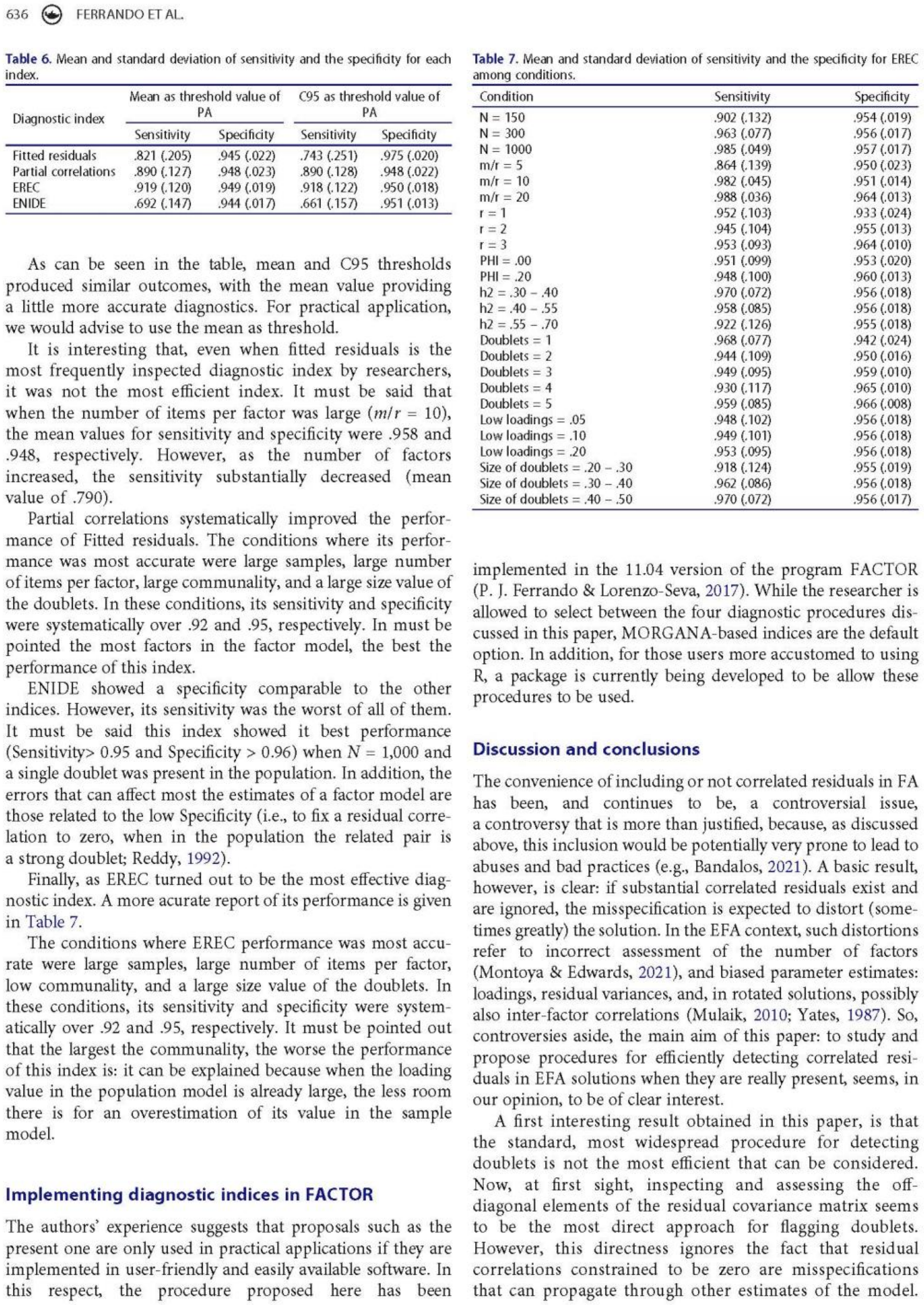


Table 6. Mean and standard deviation of sensitivity and the specificity for each index.

Diagnostic index	Mean as threshold value of PA		C95 as threshold value of PA	
	Sensitivity	Specificity	Sensitivity	Specificity
Fitted residuals	.821 (.205)	.945 (.022)	.743 (.251)	.975 (.020)
Partial correlations	.890 (.127)	.948 (.023)	.890 (.128)	.948 (.022)
EREC	.919 (.120)	.949 (.019)	.918 (.122)	.950 (.018)
ENIDE	.692 (.147)	.944 (.017)	.661 (.157)	.951 (.013)

As can be seen in the table, mean and C95 thresholds produced similar outcomes, with the mean value providing a little more accurate diagnostics. For practical application, we would advise to use the mean as threshold.

It is interesting that, even when fitted residuals is the most frequently inspected diagnostic index by researchers, it was not the most efficient index. It must be said that when the number of items per factor was large (m/r = 10), the mean values for sensitivity and specificity were .958 and .948, respectively. However, as the number of factors increased, the sensitivity substantially decreased (mean value of .790).

Partial correlations systematically improved the performance of Fitted residuals. The conditions where its performance was most accurate were large samples, large number of items per factor, large communality, and a large size value of the doublets. In these conditions, its sensitivity and specificity were systematically over .92 and .95, respectively. In must be pointed the most factors in the factor model, the best the performance of this index.

ENIDE showed a specificity comparable to the other indices. However, its sensitivity was the worst of all of them. It must be said this index showed it best performance (Sensitivity> 0.95 and Specificity > 0.96) when N = 1,000 and a single doublet was present in the population. In addition, the errors that can affect most the estimates of a factor model are those related to the low Specificity (i.e., to fix a residual correlation to zero, when in the population the related pair is a strong doublet; Reddy, 1992).

Finally, as EREC turned out to be the most effective diagnostic index. A more acurate report of its performance is given in Table 7.

The conditions where EREC performance was most accurate were large samples, large number of items per factor, low communality, and a large size value of the doublets. In these conditions, its sensitivity and specificity were systematically over .92 and .95, respectively. It must be pointed out that the largest the communality, the worse the performance of this index is: it can be explained because when the loading value in the population model is already large, the less room there is for an overestimation of its value in the sample model.

Implementing diagnostic indices in FACTOR

The authors' experience suggests that proposals such as the present one are only used in practical applications if they are implemented in user-friendly and easily available software. In this respect, the procedure proposed here has been implemented in the 11.04 version of the program FACTOR (P. J. Ferrando & Lorenzo-Seva, 2017). While the researcher is allowed to select between the four diagnostic procedures discussed in this paper, MORGANA-based indices are the default option. In addition, for those users more accustomed to using R, a package is currently being developed to be allow these procedures to be used.

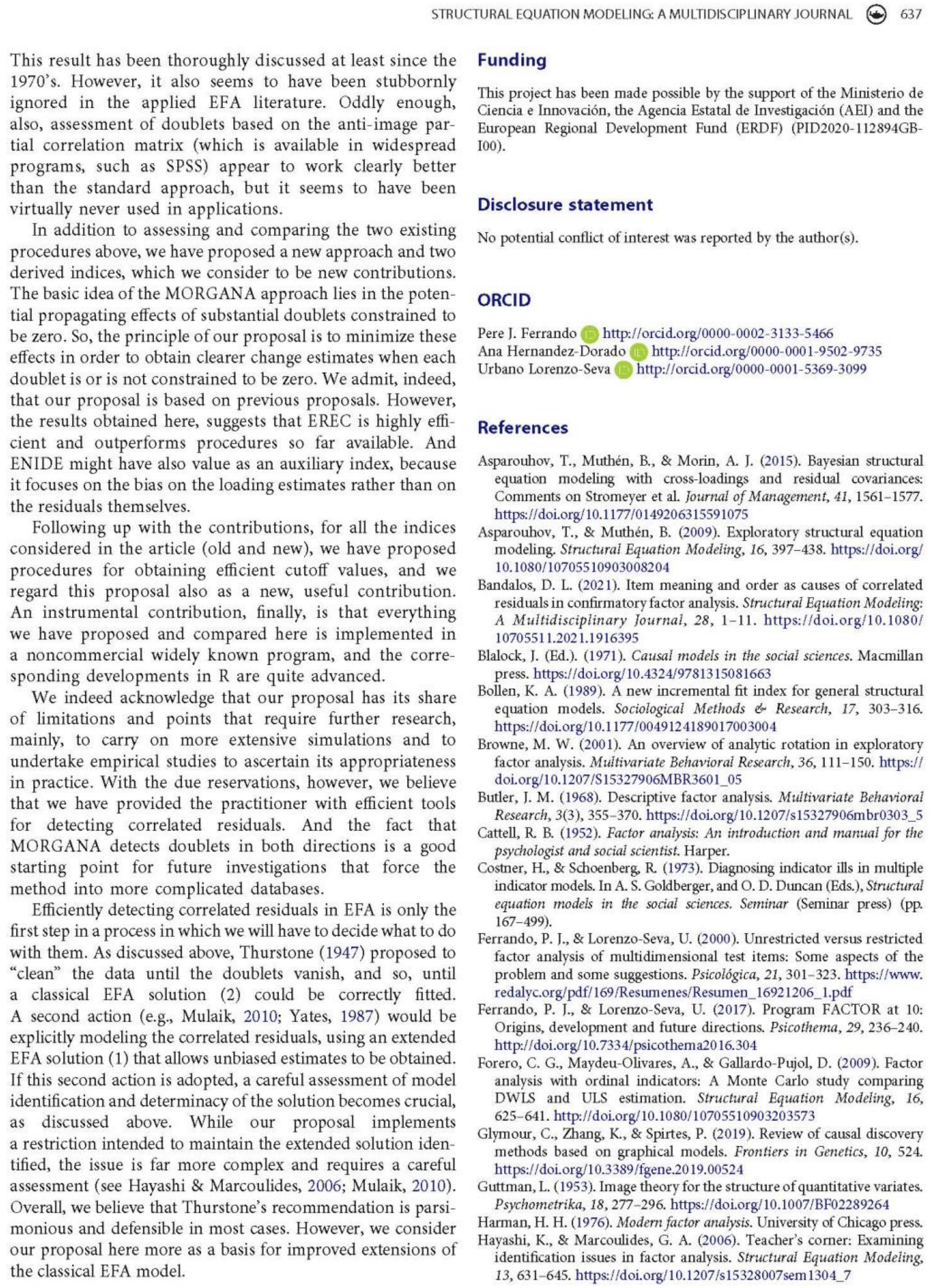
Table 7. Mean and standard deviation of sensitivity and the specificity for EREC among conditions.

Condition	Sensitivity	Specificity
N = 150	.902 (.132)	.954 (.019)
N = 300	.963 (.077)	.956 (.017)
N = 1000	.985 (.049)	.957 (.017)
m/r = 5	.864 (.139)	.950 (.023)
m/r = 10	.982 (.045)	.951 (.014)
m/r = 20	.988 (.036)	.964 (.013)
r = 1	.952 (.103)	.933 (.024)
r = 2	.945 (.104)	.955 (.013)
r = 3	.953 (.093)	.964 (.010)
PHI = .00	.951 (.099)	.953 (.020)
PHI = .20	.948 (.100)	.960 (.013)
h2 = .30 – .40	.970 (.072)	.956 (.018)
h2 = .40 – .55	.958 (.085)	.956 (.018)
h2 = .55 – .70	.922 (.126)	.955 (.018)
Doublets = 1	.968 (.077)	.942 (.024)
Doublets = 2	.944 (.109)	.950 (.016)
Doublets = 3	.949 (.095)	.959 (.010)
Doublets = 4	.930 (.117)	.965 (.010)
Doublets = 5	.959 (.085)	.966 (.008)
Low loadings = .05	.948 (.102)	.956 (.018)
Low loadings = .10	.949 (.101)	.956 (.018)
Low loadings = .20	.953 (.095)	.956 (.018)
Size of doublets = .20 – .30	.918 (.124)	.955 (.019)
Size of doublets = .30 – .40	.962 (.086)	.956 (.018)
Size of doublets = .40 – .50	.970 (.072)	.956 (.017)

Discussion and conclusions

The convenience of including or not correlated residuals in FA has been, and continues to be, a controversial issue, a controversy that is more than justified, because, as discussed above, this inclusion would be potentially very prone to lead to abuses and bad practices (e.g., Bandalos, 2021). A basic result, however, is clear: if substantial correlated residuals exist and are ignored, the misspecification is expected to distort (sometimes greatly) the solution. In the EFA context, such distortions refer to incorrect assessment of the number of factors (Montoya & Edwards, 2021), and biased parameter estimates: loadings, residual variances, and, in rotated solutions, possibly also inter-factor correlations (Mulaik, 2010; Yates, 1987). So, controversies aside, the main aim of this paper: to study and propose procedures for efficiently detecting correlated residuals in EFA solutions when they are really present, seems, in our opinion, to be of clear interest.

A first interesting result obtained in this paper, is that the standard, most widespread procedure for detecting doublets is not the most efficient that can be considered. Now, at first sight, inspecting and assessing the off-diagonal elements of the residual covariance matrix seems to be the most direct approach for flagging doublets. However, this directness ignores the fact that residual correlations constrained to be zero are misspecifications that can propagate through other estimates of the model.



This result has been thoroughly discussed at least since the 1970's. However, it also seems to have been stubbornly ignored in the applied EFA literature. Oddly enough, also, assessment of doublets based on the anti-image partial correlation matrix (which is available in widespread programs, such as SPSS) appear to work clearly better than the standard approach, but it seems to have been virtually never used in applications.

In addition to assessing and comparing the two existing procedures above, we have proposed a new approach and two derived indices, which we consider to be new contributions. The basic idea of the MORGANA approach lies in the potential propagating effects of substantial doublets constrained to be zero. So, the principle of our proposal is to minimize these effects in order to obtain clearer change estimates when each doublet is or is not constrained to be zero. We admit, indeed, that our proposal is based on previous proposals. However, the results obtained here, suggests that EREC is highly efficient and outperforms procedures so far available. And ENIDE might have also value as an auxiliary index, because it focuses on the bias on the loading estimates rather than on the residuals themselves.

Following up with the contributions, for all the indices considered in the article (old and new), we have proposed procedures for obtaining efficient cutoff values, and we regard this proposal also as a new, useful contribution. An instrumental contribution, finally, is that everything we have proposed and compared here is implemented in a noncommercial widely known program, and the corresponding developments in R are quite advanced.

We indeed acknowledge that our proposal has its share of limitations and points that require further research, mainly, to carry on more extensive simulations and to undertake empirical studies to ascertain its appropriateness in practice. With the due reservations, however, we believe that we have provided the practitioner with efficient tools for detecting correlated residuals. And the fact that MORGANA detects doublets in both directions is a good starting point for future investigations that force the method into more complicated databases.

Efficiently detecting correlated residuals in EFA is only the first step in a process in which we will have to decide what to do with them. As discussed above, Thurstone (1947) proposed to "clean" the data until the doublets vanish, and so, until a classical EFA solution (2) could be correctly fitted. A second action (e.g., Mulaik, 2010; Yates, 1987) would be explicitly modeling the correlated residuals, using an extended EFA solution (1) that allows unbiased estimates to be obtained. If this second action is adopted, a careful assessment of model identification and determinacy of the solution becomes crucial, as discussed above. While our proposal implements a restriction intended to maintain the extended solution identified, the issue is far more complex and requires a careful assessment (see Hayashi & Marcoulides, 2006; Mulaik, 2010). Overall, we believe that Thurstone's recommendation is parsimonious and defensible in most cases. However, we consider our proposal here more as a basis for improved extensions of the classical EFA model.

Funding

This project has been made possible by the support of the Ministerio de Ciencia e Innovación, the Agencia Estatal de Investigación (AEI) and the European Regional Development Fund (ERDF) (PID2020-112894GB-I00).

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Pere J. Ferrando http://orcid.org/0000-0002-3133-5466
Ana Hernandez-Dorado http://orcid.org/0000-0001-9502-9735
Urbano Lorenzo-Seva http://orcid.org/0000-0001-5369-3099

References

Asparouhov, T., Muthén, B., & Morin, A. J. (2015). Bayesian structural equation modeling with cross-loadings and residual covariances: Comments on Stromeyer et al. *Journal of Management*, *41*, 1561–1577. https://doi.org/10.1177/0149206315591075

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling*, *16*, 397–438. https://doi.org/10.1080/10705510903008204

Bandalos, D. L. (2021). Item meaning and order as causes of correlated residuals in confirmatory factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*, 1–11. https://doi.org/10.1080/10705511.2021.1916395

Blalock, J. (Ed.). (1971). *Causal models in the social sciences*. Macmillan press. https://doi.org/10.4324/9781315081663

Bollen, K. A. (1989). A new incremental fit index for general structural equation models. *Sociological Methods & Research*, *17*, 303–316. https://doi.org/10.1177/0049124189017003004

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*, *36*, 111–150. https://doi.org/10.1207/S15327906MBR3601_05

Butler, J. M. (1968). Descriptive factor analysis. *Multivariate Behavioral Research*, 3(3), 355–370. https://doi.org/10.1207/s15327906mbr0303_5

Cattell, R. B. (1952). *Factor analysis: An introduction and manual for the psychologist and social scientist*. Harper.

Costner, H., & Schoenberg, R. (1973). Diagnosing indicator ills in multiple indicator models. In A. S. Goldberger, and O. D. Duncan (Eds.), *Structural equation models in the social sciences. Seminar* (Seminar press) (pp. 167–499).

Ferrando, P. J., & Lorenzo-Seva, U. (2000). Unrestricted versus restricted factor analysis of multidimensional test items: Some aspects of the problem and some suggestions. *Psicológica*, *21*, 301–323. https://www.redalyc.org/pdf/169/Resumenes/Resumen_16921206_1.pdf

Ferrando, P. J., & Lorenzo-Seva, U. (2017). Program FACTOR at 10: Origins, development and future directions. *Psicothema*, *29*, 236–240. http://doi.org/10.7334/psicothema2016.304

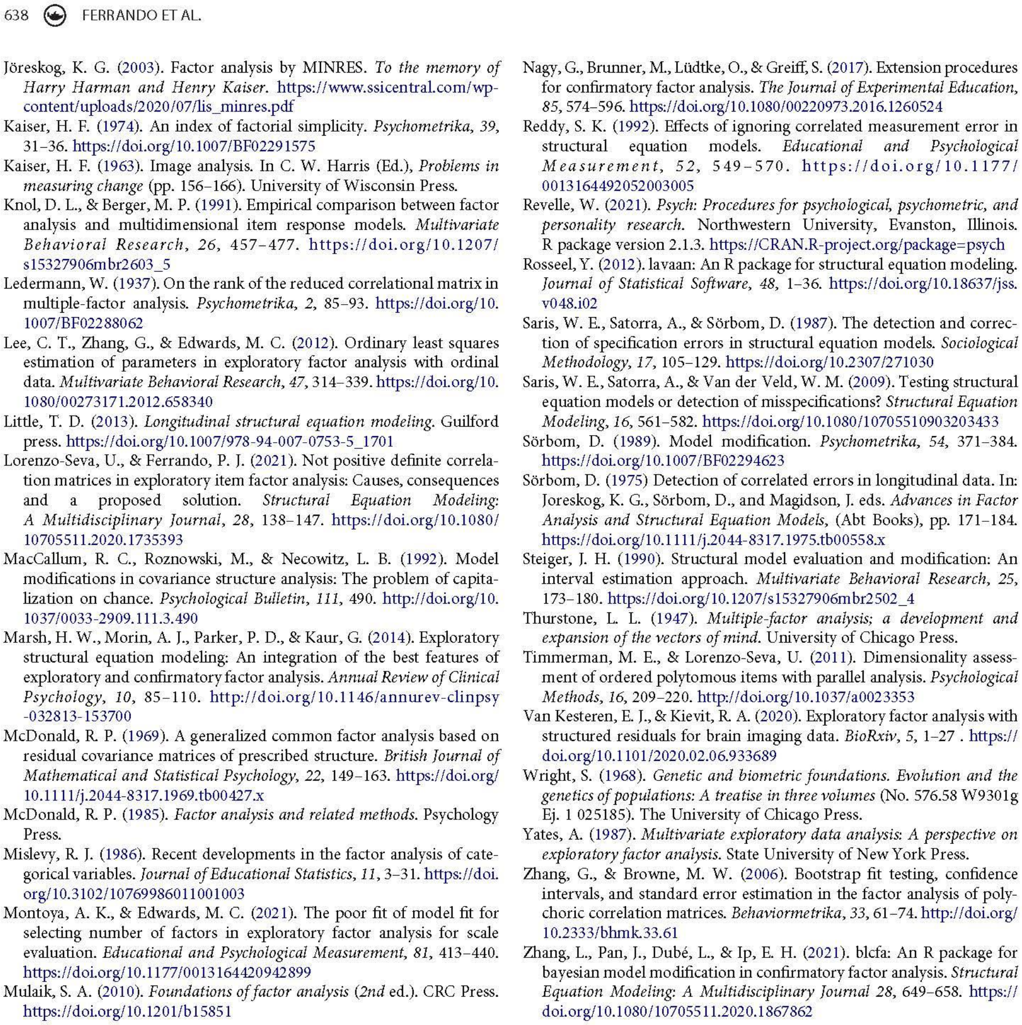
Forero, C. G., Maydeu-Olivares, A., & Gallardo-Pujol, D. (2009). Factor analysis with ordinal indicators: A Monte Carlo study comparing DWLS and ULS estimation. *Structural Equation Modeling*, *16*, 625–641. http://doi.org/10.1080/10705510903203573

Glymour, C., Zhang, K., & Spirtes, P. (2019). Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, *10*, 524. https://doi.org/10.3389/fgene.2019.00524

Guttman, L. (1953). Image theory for the structure of quantitative variates. *Psychometrika*, *18*, 277–296. https://doi.org/10.1007/BF02289264

Harman, H. H. (1976). *Modern factor analysis*. University of Chicago press.

Hayashi, K., & Marcoulides, G. A. (2006). Teacher's corner: Examining identification issues in factor analysis. *Structural Equation Modeling*, *13*, 631–645. https://doi.org/10.1207/s15328007sem1304_7



Jöreskog, K. G. (2003). Factor analysis by MINRES. *To the memory of Harry Harman and Henry Kaiser*. https://www.ssicentral.com/wp-content/uploads/2020/07/lis_minres.pdf

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika*, *39*, 31–36. https://doi.org/10.1007/BF02291575

Kaiser, H. F. (1963). Image analysis. In C. W. Harris (Ed.), *Problems in measuring change* (pp. 156–166). University of Wisconsin Press.

Knol, D. L., & Berger, M. P. (1991). Empirical comparison between factor analysis and multidimensional item response models. *Multivariate Behavioral Research*, *26*, 457–477. https://doi.org/10.1207/s15327906mbr2603_5

Ledermann, W. (1937). On the rank of the reduced correlational matrix in multiple-factor analysis. *Psychometrika*, *2*, 85–93. https://doi.org/10.1007/BF02288062

Lee, C. T., Zhang, G., & Edwards, M. C. (2012). Ordinary least squares estimation of parameters in exploratory factor analysis with ordinal data. *Multivariate Behavioral Research*, *47*, 314–339. https://doi.org/10.1080/00273171.2012.658340

Little, T. D. (2013). *Longitudinal structural equation modeling*. Guilford press. https://doi.org/10.1007/978-94-007-0753-5_1701

Lorenzo-Seva, U., & Ferrando, P. J. (2021). Not positive definite correlation matrices in exploratory item factor analysis: Causes, consequences and a proposed solution. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*, 138–147. https://doi.org/10.1080/10705511.2020.1735393

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin*, *111*, 490. http://doi.org/10.1037/0033-2909.111.3.490

Marsh, H. W., Morin, A. J., Parker, P. D., & Kaur, G. (2014). Exploratory structural equation modeling: An integration of the best features of exploratory and confirmatory factor analysis. *Annual Review of Clinical Psychology*, *10*, 85–110. http://doi.org/10.1146/annurev-clinpsy-032813-153700

McDonald, R. P. (1969). A generalized common factor analysis based on residual covariance matrices of prescribed structure. *British Journal of Mathematical and Statistical Psychology*, *22*, 149–163. https://doi.org/10.1111/j.2044-8317.1969.tb00427.x

McDonald, R. P. (1985). *Factor analysis and related methods*. Psychology Press.

Mislevy, R. J. (1986). Recent developments in the factor analysis of categorical variables. *Journal of Educational Statistics*, *11*, 3–31. https://doi.org/10.3102/10769986011001003

Montoya, A. K., & Edwards, M. C. (2021). The poor fit of model fit for selecting number of factors in exploratory factor analysis for scale evaluation. *Educational and Psychological Measurement*, *81*, 413–440. https://doi.org/10.1177/0013164420942899

Mulaik, S. A. (2010). *Foundations of factor analysis* (2nd ed.). CRC Press. https://doi.org/10.1201/b15851

Nagy, G., Brunner, M., Lüdtke, O., & Greiff, S. (2017). Extension procedures for confirmatory factor analysis. *The Journal of Experimental Education*, *85*, 574–596. https://doi.org/10.1080/00220973.2016.1260524

Reddy, S. K. (1992). Effects of ignoring correlated measurement error in structural equation models. *Educational and Psychological Measurement*, *52*, 549–570. https://doi.org/10.1177/0013164492052003005

Revelle, W. (2021). *Psych: Procedures for psychological, psychometric, and personality research*. Northwestern University, Evanston, Illinois. R package version 2.1.3. https://CRAN.R-project.org/package=psych

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*, 1–36. https://doi.org/10.18637/jss.v048.i02

Saris, W. E., Satorra, A., & Sörbom, D. (1987). The detection and correction of specification errors in structural equation models. *Sociological Methodology*, *17*, 105–129. https://doi.org/10.2307/271030

Saris, W. E., Satorra, A., & Van der Veld, W. M. (2009). Testing structural equation models or detection of misspecifications? *Structural Equation Modeling*, *16*, 561–582. https://doi.org/10.1080/10705510903203433

Sörbom, D. (1989). Model modification. *Psychometrika*, *54*, 371–384. https://doi.org/10.1007/BF02294623

Sörbom, D. (1975) Detection of correlated errors in longitudinal data. In: Joreskog, K. G., Sörbom, D., and Magidson, J. eds. *Advances in Factor Analysis and Structural Equation Models*, (Abt Books), pp. 171–184. https://doi.org/10.1111/j.2044-8317.1975.tb00558.x

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, *25*, 173–180. https://doi.org/10.1207/s15327906mbr2502_4

Thurstone, L. L. (1947). *Multiple-factor analysis; a development and expansion of the vectors of mind*. University of Chicago Press.

Timmerman, M. E., & Lorenzo-Seva, U. (2011). Dimensionality assessment of ordered polytomous items with parallel analysis. *Psychological Methods*, *16*, 209–220. http://doi.org/10.1037/a0023353

Van Kesteren, E. J., & Kievit, R. A. (2020). Exploratory factor analysis with structured residuals for brain imaging data. *BioRxiv*, *5*, 1–27 . https://doi.org/10.1101/2020.02.06.933689

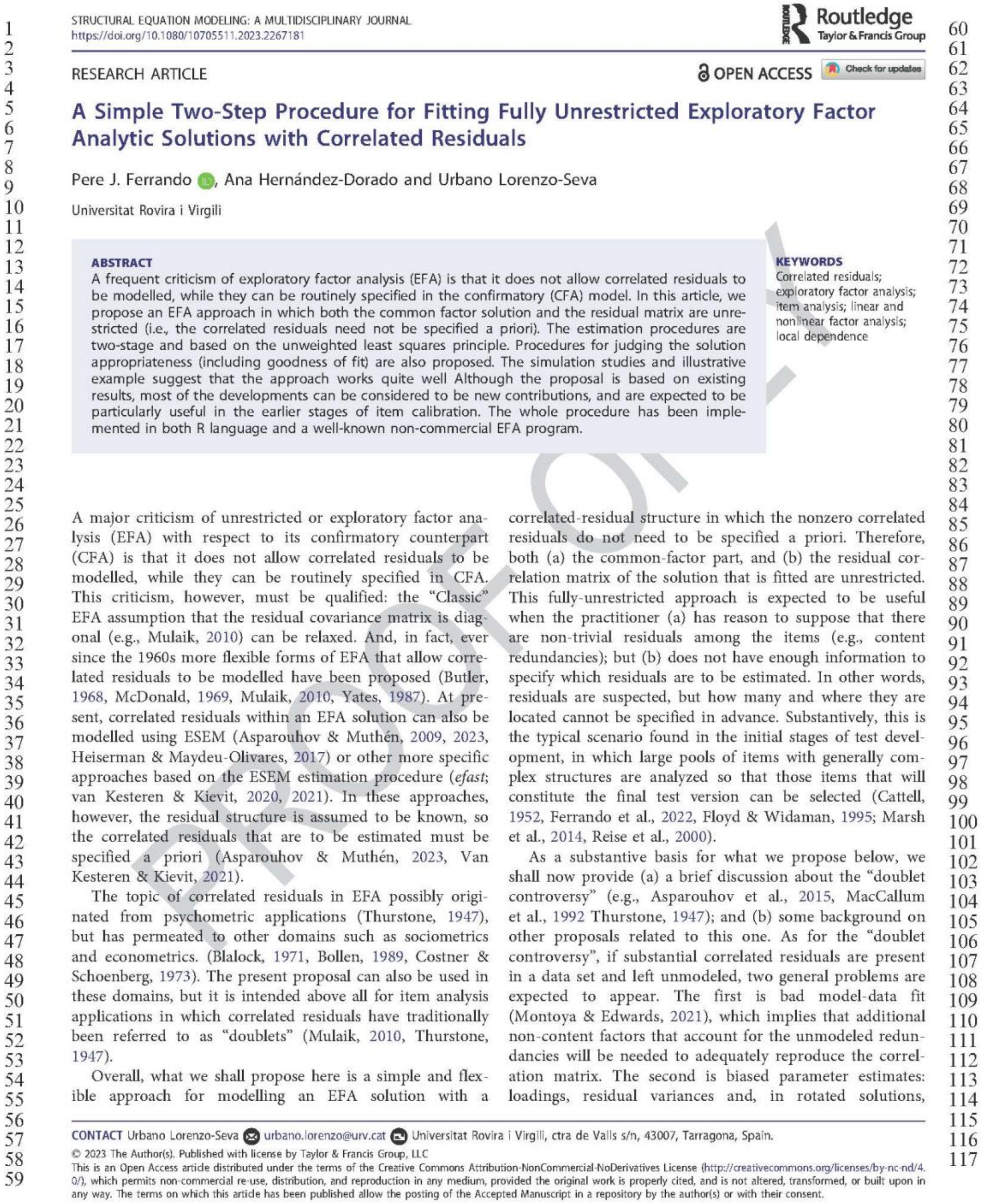
Wright, S. (1968). *Genetic and biometric foundations. Evolution and the genetics of populations: A treatise in three volumes* (No. 576.58 W9301g Ej. 1 025185). The University of Chicago Press.

Yates, A. (1987). *Multivariate exploratory data analysis: A perspective on exploratory factor analysis*. State University of New York Press.

Zhang, G., & Browne, M. W. (2006). Bootstrap fit testing, confidence intervals, and standard error estimation in the factor analysis of polychoric correlation matrices. *Behaviormetrika*, *33*, 61–74. http://doi.org/10.2333/bhmk.33.61

Zhang, L., Pan, J., Dubé, L., & Ip, E. H. (2021). blcfa: An R package for bayesian model modification in confirmatory factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal 28*, 649–658. https://doi.org/10.1080/10705511.2020.1867862

Capítulo 4. Resultados



STRUCTURAL EQUATION MODELING: A MULTIDISCIPLINARY JOURNAL
https://doi.org/10.1080/10705511.2023.2267181

Routledge
Taylor & Francis Group

RESEARCH ARTICLE

OPEN ACCESS · Check for updates

A Simple Two-Step Procedure for Fitting Fully Unrestricted Exploratory Factor Analytic Solutions with Correlated Residuals

Pere J. Ferrando, Ana Hernández-Dorado and Urbano Lorenzo-Seva

Universitat Rovira i Virgili

ABSTRACT

A frequent criticism of exploratory factor analysis (EFA) is that it does not allow correlated residuals to be modelled, while they can be routinely specified in the confirmatory (CFA) model. In this article, we propose an EFA approach in which both the common factor solution and the residual matrix are unrestricted (i.e., the correlated residuals need not be specified a priori). The estimation procedures are two-stage and based on the unweighted least squares principle. Procedures for judging the solution appropriateness (including goodness of fit) are also proposed. The simulation studies and illustrative example suggest that the approach works quite well Although the proposal is based on existing results, most of the developments can be considered to be new contributions, and are expected to be particularly useful in the earlier stages of item calibration. The whole procedure has been implemented in both R language and a well-known non-commercial EFA program.

KEYWORDS
Correlated residuals; exploratory factor analysis; item analysis; linear and nonlinear factor analysis; local dependence

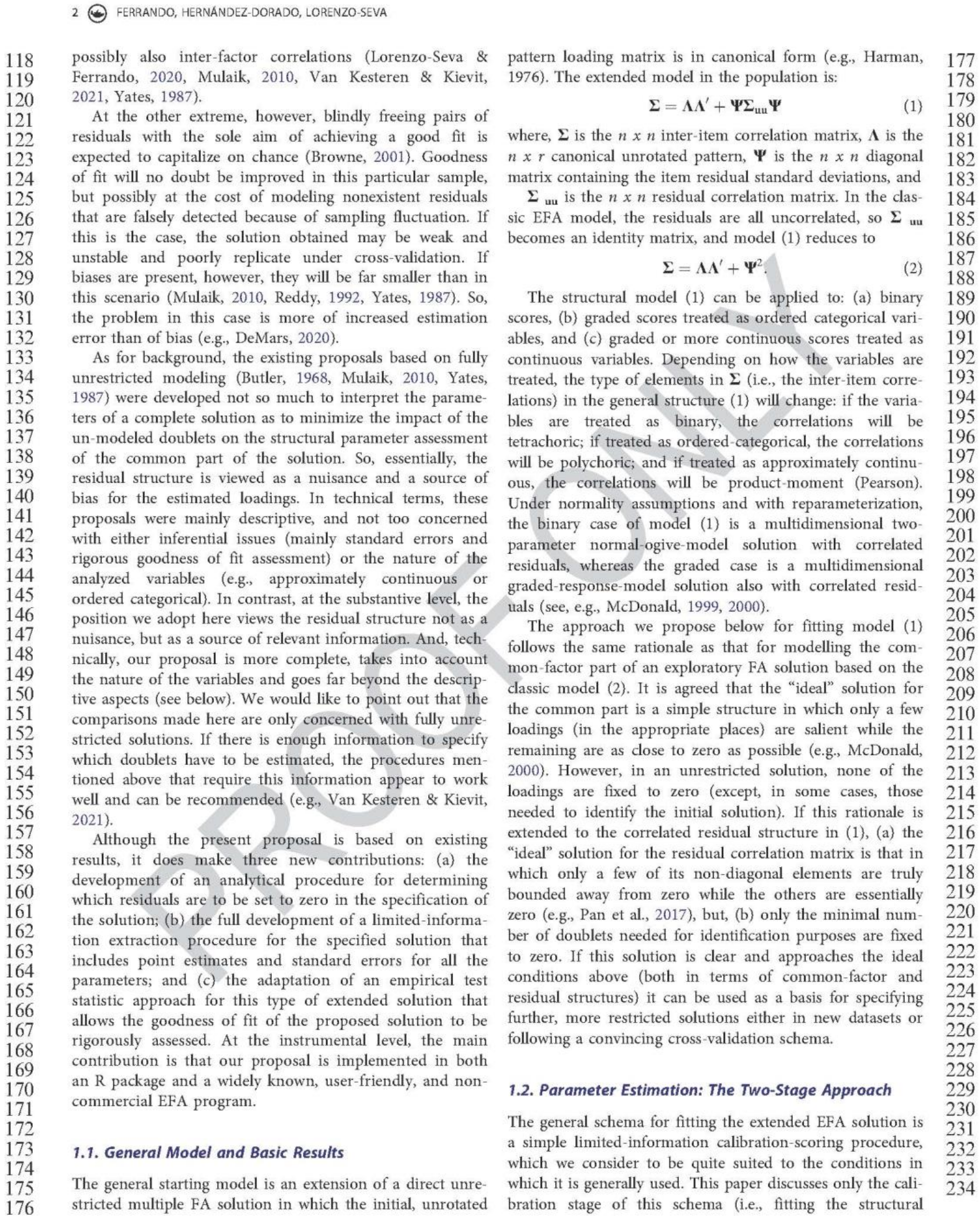
A major criticism of unrestricted or exploratory factor analysis (EFA) with respect to its confirmatory counterpart (CFA) is that it does not allow correlated residuals to be modelled, while they can be routinely specified in CFA. This criticism, however, must be qualified: the "Classic" EFA assumption that the residual covariance matrix is diagonal (e.g., Mulaik, 2010) can be relaxed. And, in fact, ever since the 1960s more flexible forms of EFA that allow correlated residuals to be modelled have been proposed (Butler, 1968, McDonald, 1969, Mulaik, 2010, Yates, 1987). At present, correlated residuals within an EFA solution can also be modelled using ESEM (Asparouhov & Muthén, 2009, 2023, Heiserman & Maydeu-Olivares, 2017) or other more specific approaches based on the ESEM estimation procedure (*efast*; van Kesteren & Kievit, 2020, 2021). In these approaches, however, the residual structure is assumed to be known, so the correlated residuals that are to be estimated must be specified a priori (Asparouhov & Muthén, 2023, Van Kesteren & Kievit, 2021).

The topic of correlated residuals in EFA possibly originated from psychometric applications (Thurstone, 1947), but has permeated to other domains such as sociometrics and econometrics. (Blalock, 1971, Bollen, 1989, Costner & Schoenberg, 1973). The present proposal can also be used in these domains, but it is intended above all for item analysis applications in which correlated residuals have traditionally been referred to as "doublets" (Mulaik, 2010, Thurstone, 1947).

Overall, what we shall propose here is a simple and flexible approach for modelling an EFA solution with a correlated-residual structure in which the nonzero correlated residuals do not need to be specified a priori. Therefore, both (a) the common-factor part, and (b) the residual correlation matrix of the solution that is fitted are unrestricted. This fully-unrestricted approach is expected to be useful when the practitioner (a) has reason to suppose that there are non-trivial residuals among the items (e.g., content redundancies); but (b) does not have enough information to specify which residuals are to be estimated. In other words, residuals are suspected, but how many and where they are located cannot be specified in advance. Substantively, this is the typical scenario found in the initial stages of test development, in which large pools of items with generally complex structures are analyzed so that those items that will constitute the final test version can be selected (Cattell, 1952, Ferrando et al., 2022, Floyd & Widaman, 1995; Marsh et al., 2014, Reise et al., 2000).

As a substantive basis for what we propose below, we shall now provide (a) a brief discussion about the "doublet controversy" (e.g., Asparouhov et al., 2015, MacCallum et al., 1992 Thurstone, 1947); and (b) some background on other proposals related to this one. As for the "doublet controversy", if substantial correlated residuals are present in a data set and left unmodeled, two general problems are expected to appear. The first is bad model-data fit (Montoya & Edwards, 2021), which implies that additional non-content factors that account for the unmodeled redundancies will be needed to adequately reproduce the correlation matrix. The second is biased parameter estimates: loadings, residual variances and, in rotated solutions,

CONTACT Urbano Lorenzo-Seva urbano.lorenzo@urv.cat Universitat Rovira i Virgili, ctra de Valls s/n, 43007, Tarragona, Spain.



possibly also inter-factor correlations (Lorenzo-Seva & Ferrando, 2020, Mulaik, 2010, Van Kesteren & Kievit, 2021, Yates, 1987).

At the other extreme, however, blindly freeing pairs of residuals with the sole aim of achieving a good fit is expected to capitalize on chance (Browne, 2001). Goodness of fit will no doubt be improved in this particular sample, but possibly at the cost of modeling nonexistent residuals that are falsely detected because of sampling fluctuation. If this is the case, the solution obtained may be weak and unstable and poorly replicate under cross-validation. If biases are present, however, they will be far smaller than in this scenario (Mulaik, 2010, Reddy, 1992, Yates, 1987). So, the problem in this case is more of increased estimation error than of bias (e.g., DeMars, 2020).

As for background, the existing proposals based on fully unrestricted modeling (Butler, 1968, Mulaik, 2010, Yates, 1987) were developed not so much to interpret the parameters of a complete solution as to minimize the impact of the un-modeled doublets on the structural parameter assessment of the common part of the solution. So, essentially, the residual structure is viewed as a nuisance and a source of bias for the estimated loadings. In technical terms, these proposals were mainly descriptive, and not too concerned with either inferential issues (mainly standard errors and rigorous goodness of fit assessment) or the nature of the analyzed variables (e.g., approximately continuous or ordered categorical). In contrast, at the substantive level, the position we adopt here views the residual structure not as a nuisance, but as a source of relevant information. And, technically, our proposal is more complete, takes into account the nature of the variables and goes far beyond the descriptive aspects (see below). We would like to point out that the comparisons made here are only concerned with fully unrestricted solutions. If there is enough information to specify which doublets have to be estimated, the procedures mentioned above that require this information appear to work well and can be recommended (e.g., Van Kesteren & Kievit, 2021).

Although the present proposal is based on existing results, it does make three new contributions: (a) the development of an analytical procedure for determining which residuals are to be set to zero in the specification of the solution; (b) the full development of a limited-information extraction procedure for the specified solution that includes point estimates and standard errors for all the parameters; and (c) the adaptation of an empirical test statistic approach for this type of extended solution that allows the goodness of fit of the proposed solution to be rigorously assessed. At the instrumental level, the main contribution is that our proposal is implemented in both an R package and a widely known, user-friendly, and non-commercial EFA program.

1.1. General Model and Basic Results

The general starting model is an extension of a direct unrestricted multiple FA solution in which the initial, unrotated pattern loading matrix is in canonical form (e.g., Harman, 1976). The extended model in the population is:

$$\Sigma = \Lambda\Lambda' + \Psi\Sigma_{uu}\Psi \qquad (1)$$

where, Σ is the n x n inter-item correlation matrix, Λ is the n x r canonical unrotated pattern, Ψ is the n x n diagonal matrix containing the item residual standard deviations, and Σ_{uu} is the n x n residual correlation matrix. In the classic EFA model, the residuals are all uncorrelated, so Σ_{uu} becomes an identity matrix, and model (1) reduces to

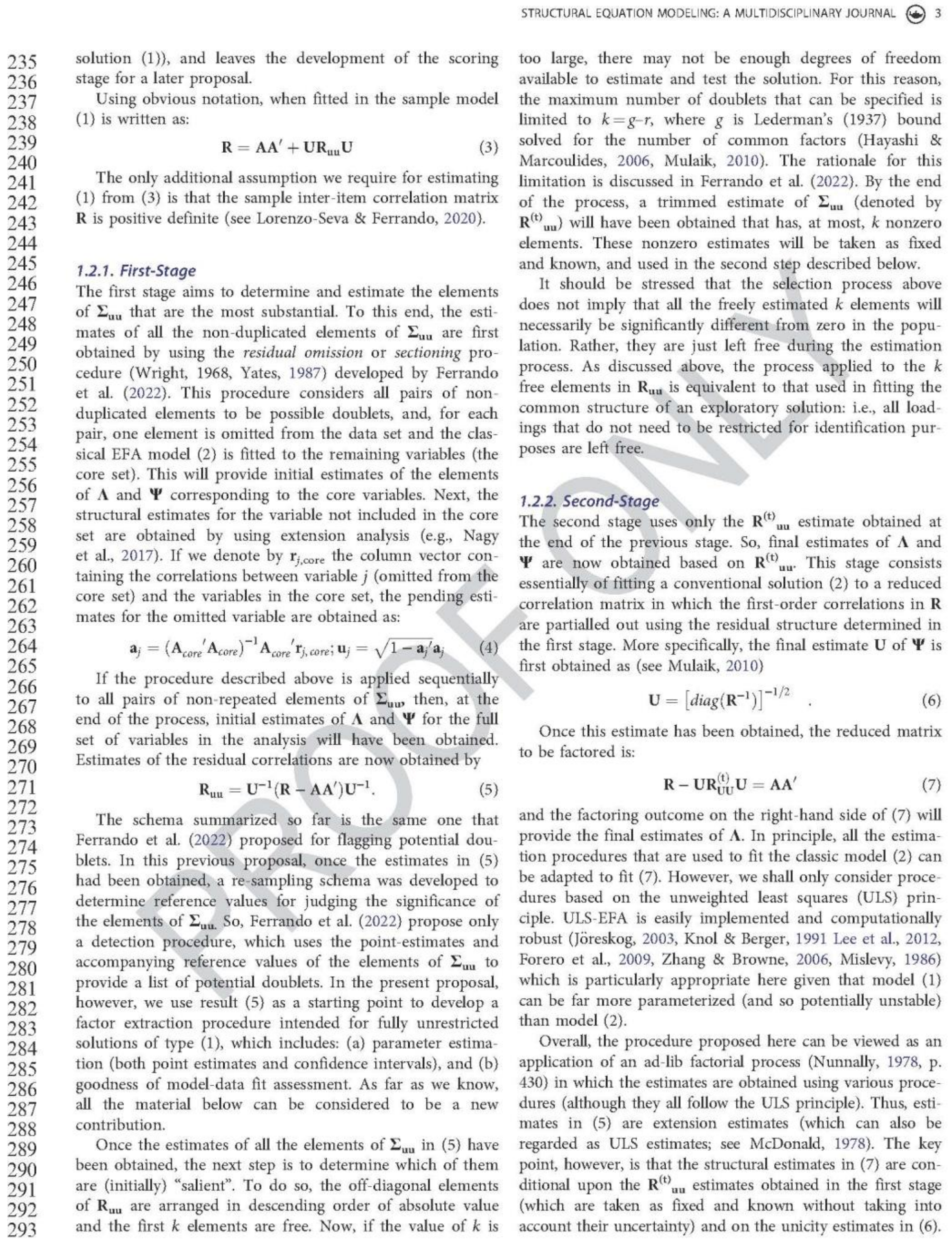
$$\Sigma = \Lambda\Lambda' + \Psi^2. \qquad (2)$$

The structural model (1) can be applied to: (a) binary scores, (b) graded scores treated as ordered categorical variables, and (c) graded or more continuous scores treated as continuous variables. Depending on how the variables are treated, the type of elements in Σ (i.e., the inter-item correlations) in the general structure (1) will change: if the variables are treated as binary, the correlations will be tetrachoric; if treated as ordered-categorical, the correlations will be polychoric; and if treated as approximately continuous, the correlations will be product-moment (Pearson). Under normality assumptions and with reparameterization, the binary case of model (1) is a multidimensional two-parameter normal-ogive-model solution with correlated residuals, whereas the graded case is a multidimensional graded-response-model solution also with correlated residuals (see, e.g., McDonald, 1999, 2000).

The approach we propose below for fitting model (1) follows the same rationale as that for modelling the common-factor part of an exploratory FA solution based on the classic model (2). It is agreed that the "ideal" solution for the common part is a simple structure in which only a few loadings (in the appropriate places) are salient while the remaining are as close to zero as possible (e.g., McDonald, 2000). However, in an unrestricted solution, none of the loadings are fixed to zero (except, in some cases, those needed to identify the initial solution). If this rationale is extended to the correlated residual structure in (1), (a) the "ideal" solution for the residual correlation matrix is that in which only a few of its non-diagonal elements are truly bounded away from zero while the others are essentially zero (e.g., Pan et al., 2017), but, (b) only the minimal number of doublets needed for identification purposes are fixed to zero. If this solution is clear and approaches the ideal conditions above (both in terms of common-factor and residual structures) it can be used as a basis for specifying further, more restricted solutions either in new datasets or following a convincing cross-validation schema.

1.2. Parameter Estimation: The Two-Stage Approach

The general schema for fitting the extended EFA solution is a simple limited-information calibration-scoring procedure, which we consider to be quite suited to the conditions in which it is generally used. This paper discusses only the calibration stage of this schema (i.e., fitting the structural



solution (1)), and leaves the development of the scoring stage for a later proposal.

Using obvious notation, when fitted in the sample model (1) is written as:

$$\mathbf{R} = \mathbf{A}\mathbf{A}' + \mathbf{U}\mathbf{R}_{\mathbf{uu}}\mathbf{U} \tag{3}$$

The only additional assumption we require for estimating (1) from (3) is that the sample inter-item correlation matrix **R** is positive definite (see Lorenzo-Seva & Ferrando, 2020).

1.2.1. First-Stage

The first stage aims to determine and estimate the elements of $\mathbf{\Sigma}_{\mathbf{uu}}$ that are the most substantial. To this end, the estimates of all the non-duplicated elements of $\mathbf{\Sigma}_{\mathbf{uu}}$ are first obtained by using the *residual omission* or *sectioning* procedure (Wright, 1968, Yates, 1987) developed by Ferrando et al. (2022). This procedure considers all pairs of non-duplicated elements to be possible doublets, and, for each pair, one element is omitted from the data set and the classical EFA model (2) is fitted to the remaining variables (the core set). This will provide initial estimates of the elements of $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ corresponding to the core variables. Next, the structural estimates for the variable not included in the core set are obtained by using extension analysis (e.g., Nagy et al., 2017). If we denote by $\mathbf{r}_{j,\text{core}}$ the column vector containing the correlations between variable j (omitted from the core set) and the variables in the core set, the pending estimates for the omitted variable are obtained as:

$$\mathbf{a}_j = (\mathbf{A}_{core}{}'\mathbf{A}_{core})^{-1}\mathbf{A}_{core}{}'\mathbf{r}_{j,core}; \mathbf{u}_j = \sqrt{1 - \mathbf{a}_j{}'\mathbf{a}_j} \tag{4}$$

If the procedure described above is applied sequentially to all pairs of non-repeated elements of $\mathbf{\Sigma}_{\mathbf{uu}}$, then, at the end of the process, initial estimates of $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ for the full set of variables in the analysis will have been obtained. Estimates of the residual correlations are now obtained by

$$\mathbf{R}_{\mathbf{uu}} = \mathbf{U}^{-1}(\mathbf{R} - \mathbf{A}\mathbf{A}')\mathbf{U}^{-1}. \tag{5}$$

The schema summarized so far is the same one that Ferrando et al. (2022) proposed for flagging potential doublets. In this previous proposal, once the estimates in (5) had been obtained, a re-sampling schema was developed to determine reference values for judging the significance of the elements of $\mathbf{\Sigma}_{\mathbf{uu}}$. So, Ferrando et al. (2022) propose only a detection procedure, which uses the point-estimates and accompanying reference values of the elements of $\mathbf{\Sigma}_{\mathbf{uu}}$ to provide a list of potential doublets. In the present proposal, however, we use result (5) as a starting point to develop a factor extraction procedure intended for fully unrestricted solutions of type (1), which includes: (a) parameter estimation (both point estimates and confidence intervals), and (b) goodness of model-data fit assessment. As far as we know, all the material below can be considered to be a new contribution.

Once the estimates of all the elements of $\mathbf{\Sigma}_{\mathbf{uu}}$ in (5) have been obtained, the next step is to determine which of them are (initially) "salient". To do so, the off-diagonal elements of $\mathbf{R}_{\mathbf{uu}}$ are arranged in descending order of absolute value and the first k elements are free. Now, if the value of k is too large, there may not be enough degrees of freedom available to estimate and test the solution. For this reason, the maximum number of doublets that can be specified is limited to $k = g–r$, where g is Lederman's (1937) bound solved for the number of common factors (Hayashi & Marcoulides, 2006, Mulaik, 2010). The rationale for this limitation is discussed in Ferrando et al. (2022). By the end of the process, a trimmed estimate of $\mathbf{\Sigma}_{\mathbf{uu}}$ (denoted by $\mathbf{R}^{(t)}{}_{\mathbf{uu}}$) will have been obtained that has, at most, k nonzero elements. These nonzero estimates will be taken as fixed and known, and used in the second step described below.

It should be stressed that the selection process above does not imply that all the freely estimated k elements will necessarily be significantly different from zero in the population. Rather, they are just left free during the estimation process. As discussed above, the process applied to the k free elements in $\mathbf{R}_{\mathbf{uu}}$ is equivalent to that used in fitting the common structure of an exploratory solution: i.e., all loadings that do not need to be restricted for identification purposes are left free.

1.2.2. Second-Stage

The second stage uses only the $\mathbf{R}^{(t)}{}_{\mathbf{uu}}$ estimate obtained at the end of the previous stage. So, final estimates of $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ are now obtained based on $\mathbf{R}^{(t)}{}_{\mathbf{uu}}$. This stage consists essentially of fitting a conventional solution (2) to a reduced correlation matrix in which the first-order correlations in **R** are partialled out using the residual structure determined in the first stage. More specifically, the final estimate **U** of $\mathbf{\Psi}$ is first obtained as (see Mulaik, 2010)

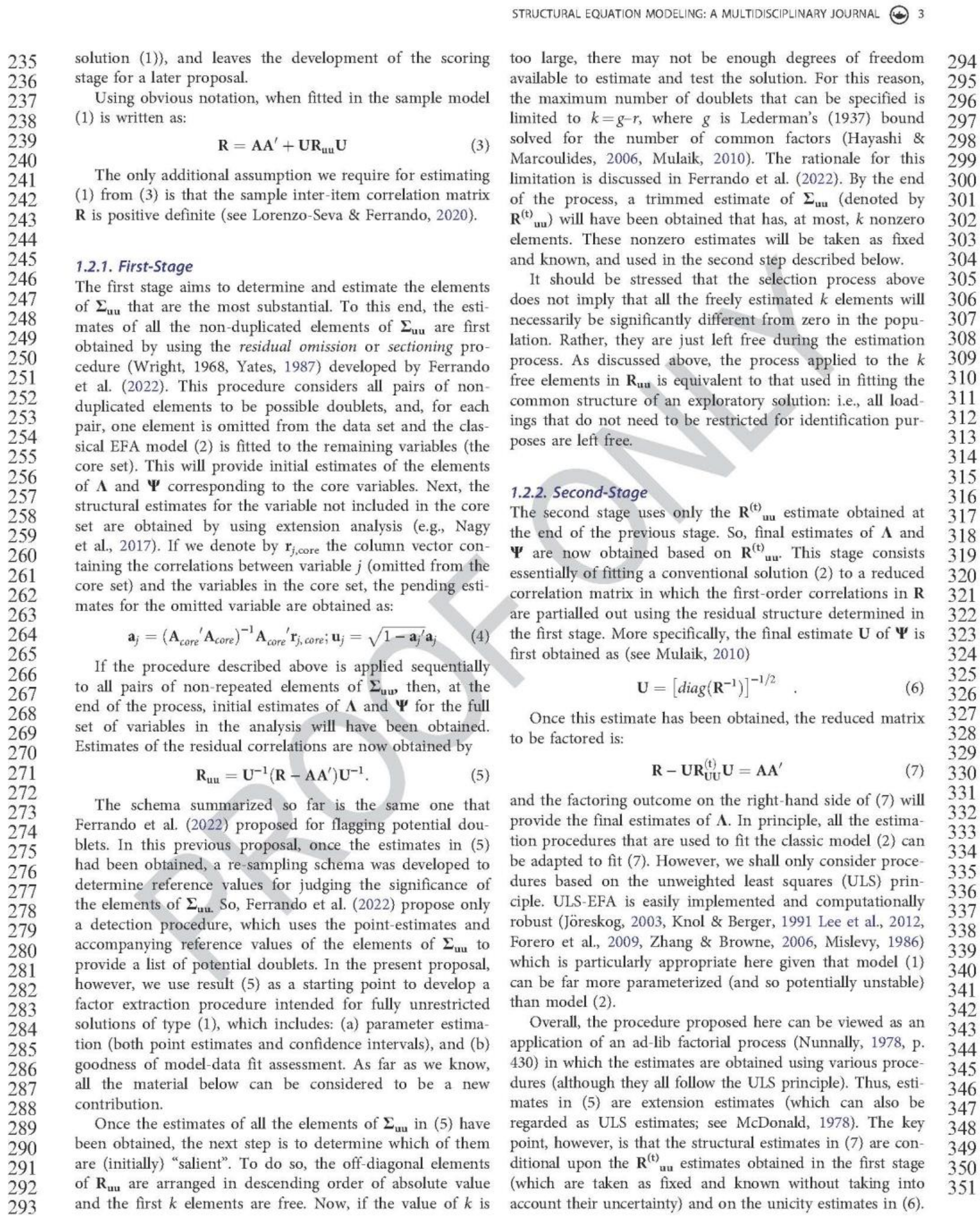
$$\mathbf{U} = \left[diag(\mathbf{R}^{-1})\right]^{-1/2} \quad . \tag{6}$$

Once this estimate has been obtained, the reduced matrix to be factored is:

$$\mathbf{R} - \mathbf{U}\mathbf{R}^{(t)}_{\mathbf{UU}}\mathbf{U} = \mathbf{A}\mathbf{A}' \tag{7}$$

and the factoring outcome on the right-hand side of (7) will provide the final estimates of $\mathbf{\Lambda}$. In principle, all the estimation procedures that are used to fit the classic model (2) can be adapted to fit (7). However, we shall only consider procedures based on the unweighted least squares (ULS) principle. ULS-EFA is easily implemented and computationally robust (Jöreskog, 2003, Knol & Berger, 1991 Lee et al., 2012, Forero et al., 2009, Zhang & Browne, 2006, Mislevy, 1986) which is particularly appropriate here given that model (1) can be far more parameterized (and so potentially unstable) than model (2).

Overall, the procedure proposed here can be viewed as an application of an ad-lib factorial process (Nunnally, 1978, p. 430) in which the estimates are obtained using various procedures (although they all follow the ULS principle). Thus, estimates in (5) are extension estimates (which can also be regarded as ULS estimates; see McDonald, 1978). The key point, however, is that the structural estimates in (7) are conditional upon the $\mathbf{R}^{(t)}{}_{\mathbf{uu}}$ estimates obtained in the first stage (which are taken as fixed and known without taking into account their uncertainty) and on the unicity estimates in (6).



solution (1)), and leaves the development of the scoring stage for a later proposal.

Using obvious notation, when fitted in the sample model (1) is written as:

$$\mathbf{R} = \mathbf{A}\mathbf{A}' + \mathbf{U}\mathbf{R}_{\mathbf{uu}}\mathbf{U} \qquad (3)$$

The only additional assumption we require for estimating (1) from (3) is that the sample inter-item correlation matrix **R** is positive definite (see Lorenzo-Seva & Ferrando, 2020).

1.2.1. First-Stage

The first stage aims to determine and estimate the elements of $\mathbf{\Sigma}_{\mathbf{uu}}$ that are the most substantial. To this end, the estimates of all the non-duplicated elements of $\mathbf{\Sigma}_{\mathbf{uu}}$ are first obtained by using the *residual omission* or *sectioning* procedure (Wright, 1968, Yates, 1987) developed by Ferrando et al. (2022). This procedure considers all pairs of non-duplicated elements to be possible doublets, and, for each pair, one element is omitted from the data set and the classical EFA model (2) is fitted to the remaining variables (the core set). This will provide initial estimates of the elements of $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ corresponding to the core variables. Next, the structural estimates for the variable not included in the core set are obtained by using extension analysis (e.g., Nagy et al., 2017). If we denote by $\mathbf{r}_{j,core}$ the column vector containing the correlations between variable *j* (omitted from the core set) and the variables in the core set, the pending estimates for the omitted variable are obtained as:

$$\mathbf{a}_j = (\mathbf{A}_{core}{}'\mathbf{A}_{core})^{-1}\mathbf{A}_{core}{}'\mathbf{r}_{j,core};\ \mathbf{u}_j = \sqrt{1-\mathbf{a}_j{}'\mathbf{a}_j} \qquad (4)$$

If the procedure described above is applied sequentially to all pairs of non-repeated elements of $\mathbf{\Sigma}_{\mathbf{uu}}$, then, at the end of the process, initial estimates of $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ for the full set of variables in the analysis will have been obtained. Estimates of the residual correlations are now obtained by

$$\mathbf{R}_{\mathbf{uu}} = \mathbf{U}^{-1}(\mathbf{R} - \mathbf{A}\mathbf{A}')\mathbf{U}^{-1}. \qquad (5)$$

The schema summarized so far is the same one that Ferrando et al. (2022) proposed for flagging potential doublets. In this previous proposal, once the estimates in (5) had been obtained, a re-sampling schema was developed to determine reference values for judging the significance of the elements of $\mathbf{\Sigma}_{\mathbf{uu}}$. So, Ferrando et al. (2022) propose only a detection procedure, which uses the point-estimates and accompanying reference values of the elements of $\mathbf{\Sigma}_{\mathbf{uu}}$ to provide a list of potential doublets. In the present proposal, however, we use result (5) as a starting point to develop a factor extraction procedure intended for fully unrestricted solutions of type (1), which includes: (a) parameter estimation (both point estimates and confidence intervals), and (b) goodness of model-data fit assessment. As far as we know, all the material below can be considered to be a new contribution.

Once the estimates of all the elements of $\mathbf{\Sigma}_{\mathbf{uu}}$ in (5) have been obtained, the next step is to determine which of them are (initially) "salient". To do so, the off-diagonal elements of $\mathbf{R}_{\mathbf{uu}}$ are arranged in descending order of absolute value and the first *k* elements are free. Now, if the value of *k* is too large, there may not be enough degrees of freedom available to estimate and test the solution. For this reason, the maximum number of doublets that can be specified is limited to $k = g - r$, where *g* is Lederman's (1937) bound solved for the number of common factors (Hayashi & Marcoulides, 2006, Mulaik, 2010). The rationale for this limitation is discussed in Ferrando et al. (2022). By the end of the process, a trimmed estimate of $\mathbf{\Sigma}_{\mathbf{uu}}$ (denoted by $\mathbf{R}^{(t)}{}_{\mathbf{uu}}$) will have been obtained that has, at most, *k* nonzero elements. These nonzero estimates will be taken as fixed and known, and used in the second step described below.

It should be stressed that the selection process above does not imply that all the freely estimated *k* elements will necessarily be significantly different from zero in the population. Rather, they are just left free during the estimation process. As discussed above, the process applied to the *k* free elements in $\mathbf{R}_{\mathbf{uu}}$ is equivalent to that used in fitting the common structure of an exploratory solution: i.e., all loadings that do not need to be restricted for identification purposes are left free.

1.2.2. Second-Stage

The second stage uses only the $\mathbf{R}^{(t)}{}_{\mathbf{uu}}$ estimate obtained at the end of the previous stage. So, final estimates of $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ are now obtained based on $\mathbf{R}^{(t)}{}_{\mathbf{uu}}$. This stage consists essentially of fitting a conventional solution (2) to a reduced correlation matrix in which the first-order correlations in **R** are partialled out using the residual structure determined in the first stage. More specifically, the final estimate **U** of $\mathbf{\Psi}$ is first obtained as (see Mulaik, 2010)

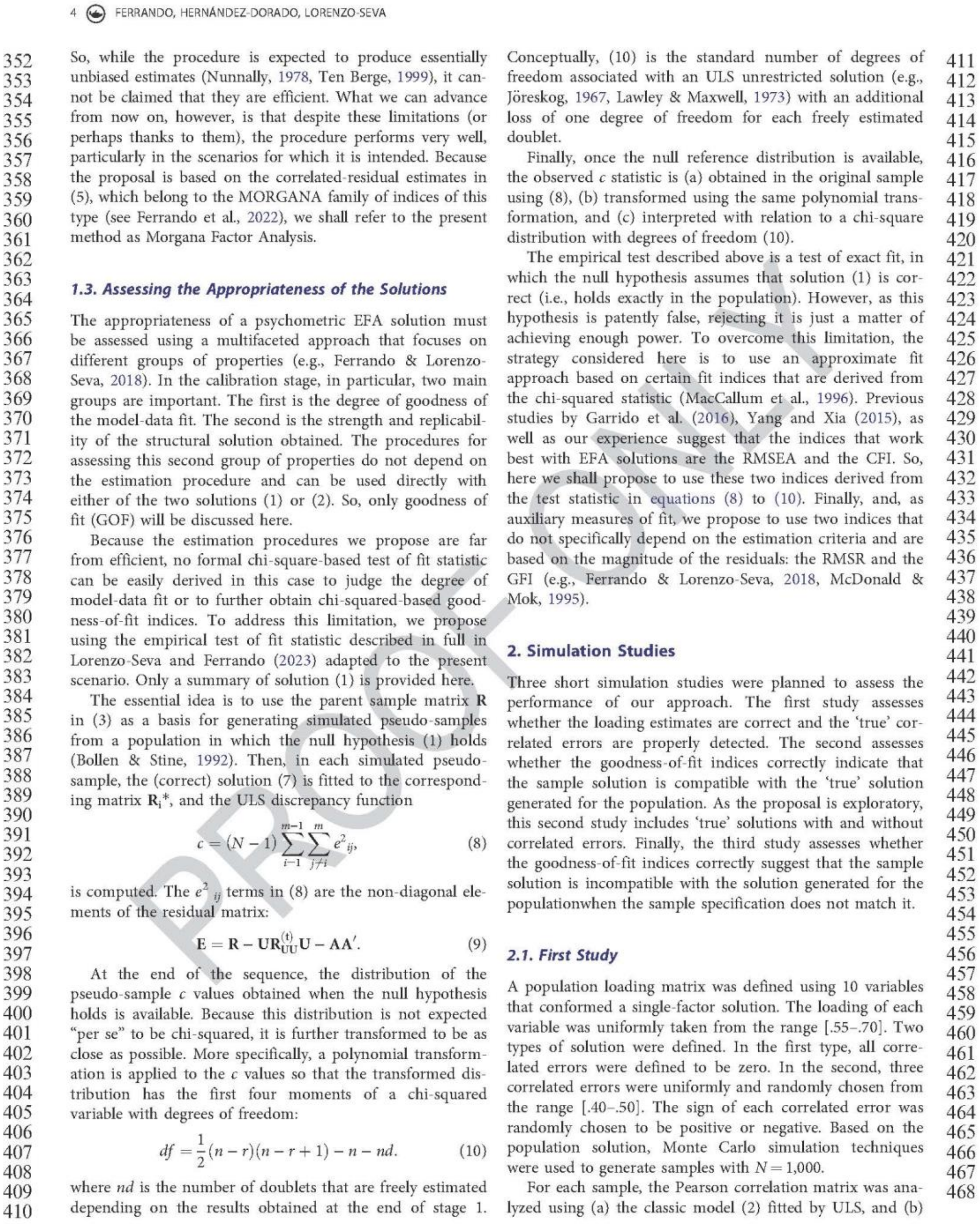
$$\mathbf{U} = \left[diag(\mathbf{R}^{-1})\right]^{-1/2} \quad . \qquad (6)$$

Once this estimate has been obtained, the reduced matrix to be factored is:

$$\mathbf{R} - \mathbf{U}\mathbf{R}^{(t)}_{\mathbf{UU}}\mathbf{U} = \mathbf{A}\mathbf{A}' \qquad (7)$$

and the factoring outcome on the right-hand side of (7) will provide the final estimates of $\mathbf{\Lambda}$. In principle, all the estimation procedures that are used to fit the classic model (2) can be adapted to fit (7). However, we shall only consider procedures based on the unweighted least squares (ULS) principle. ULS-EFA is easily implemented and computationally robust (Jöreskog, 2003, Knol & Berger, 1991 Lee et al., 2012, Forero et al., 2009, Zhang & Browne, 2006, Mislevy, 1986) which is particularly appropriate here given that model (1) can be far more parameterized (and so potentially unstable) than model (2).

Overall, the procedure proposed here can be viewed as an application of an ad-lib factorial process (Nunnally, 1978, p. 430) in which the estimates are obtained using various procedures (although they all follow the ULS principle). Thus, estimates in (5) are extension estimates (which can also be regarded as ULS estimates; see McDonald, 1978). The key point, however, is that the structural estimates in (7) are conditional upon the $\mathbf{R}^{(t)}{}_{\mathbf{uu}}$ estimates obtained in the first stage (which are taken as fixed and known without taking into account their uncertainty) and on the unicity estimates in (6).



So, while the procedure is expected to produce essentially unbiased estimates (Nunnally, 1978, Ten Berge, 1999), it cannot be claimed that they are efficient. What we can advance from now on, however, is that despite these limitations (or perhaps thanks to them), the procedure performs very well, particularly in the scenarios for which it is intended. Because the proposal is based on the correlated-residual estimates in (5), which belong to the MORGANA family of indices of this type (see Ferrando et al., 2022), we shall refer to the present method as Morgana Factor Analysis.

1.3. Assessing the Appropriateness of the Solutions

The appropriateness of a psychometric EFA solution must be assessed using a multifaceted approach that focuses on different groups of properties (e.g., Ferrando & Lorenzo-Seva, 2018). In the calibration stage, in particular, two main groups are important. The first is the degree of goodness of the model-data fit. The second is the strength and replicability of the structural solution obtained. The procedures for assessing this second group of properties do not depend on the estimation procedure and can be used directly with either of the two solutions (1) or (2). So, only goodness of fit (GOF) will be discussed here.

Because the estimation procedures we propose are far from efficient, no formal chi-square-based test of fit statistic can be easily derived in this case to judge the degree of model-data fit or to further obtain chi-squared-based goodness-of-fit indices. To address this limitation, we propose using the empirical test of fit statistic described in full in Lorenzo-Seva and Ferrando (2023) adapted to the present scenario. Only a summary of solution (1) is provided here.

The essential idea is to use the parent sample matrix **R** in (3) as a basis for generating simulated pseudo-samples from a population in which the null hypothesis (1) holds (Bollen & Stine, 1992). Then, in each simulated pseudo-sample, the (correct) solution (7) is fitted to the corresponding matrix $\mathbf{R}_i$*, and the ULS discrepancy function

$$c = (N-1)\sum_{i-1}^{m-1}\sum_{j\neq i}^{m} e^2_{ij}, \qquad (8)$$

is computed. The $e^2{}_{ij}$ terms in (8) are the non-diagonal elements of the residual matrix:

$$\mathbf{E} = \mathbf{R} - \mathbf{U}\mathbf{R}^{(t)}_{\mathrm{UU}}\mathbf{U} - \mathbf{A}\mathbf{A}'. \qquad (9)$$

At the end of the sequence, the distribution of the pseudo-sample c values obtained when the null hypothesis holds is available. Because this distribution is not expected "per se" to be chi-squared, it is further transformed to be as close as possible. More specifically, a polynomial transformation is applied to the c values so that the transformed distribution has the first four moments of a chi-squared variable with degrees of freedom:

$$df = \frac{1}{2}(n-r)(n-r+1) - n - nd. \qquad (10)$$

where nd is the number of doublets that are freely estimated depending on the results obtained at the end of stage 1. Conceptually, (10) is the standard number of degrees of freedom associated with an ULS unrestricted solution (e.g., Jöreskog, 1967, Lawley & Maxwell, 1973) with an additional loss of one degree of freedom for each freely estimated doublet.

Finally, once the null reference distribution is available, the observed c statistic is (a) obtained in the original sample using (8), (b) transformed using the same polynomial transformation, and (c) interpreted with relation to a chi-square distribution with degrees of freedom (10).

The empirical test described above is a test of exact fit, in which the null hypothesis assumes that solution (1) is correct (i.e., holds exactly in the population). However, as this hypothesis is patently false, rejecting it is just a matter of achieving enough power. To overcome this limitation, the strategy considered here is to use an approximate fit approach based on certain fit indices that are derived from the chi-squared statistic (MacCallum et al., 1996). Previous studies by Garrido et al. (2016), Yang and Xia (2015), as well as our experience suggest that the indices that work best with EFA solutions are the RMSEA and the CFI. So, here we shall propose to use these two indices derived from the test statistic in equations (8) to (10). Finally, and, as auxiliary measures of fit, we propose to use two indices that do not specifically depend on the estimation criteria and are based on the magnitude of the residuals: the RMSR and the GFI (e.g., Ferrando & Lorenzo-Seva, 2018, McDonald & Mok, 1995).

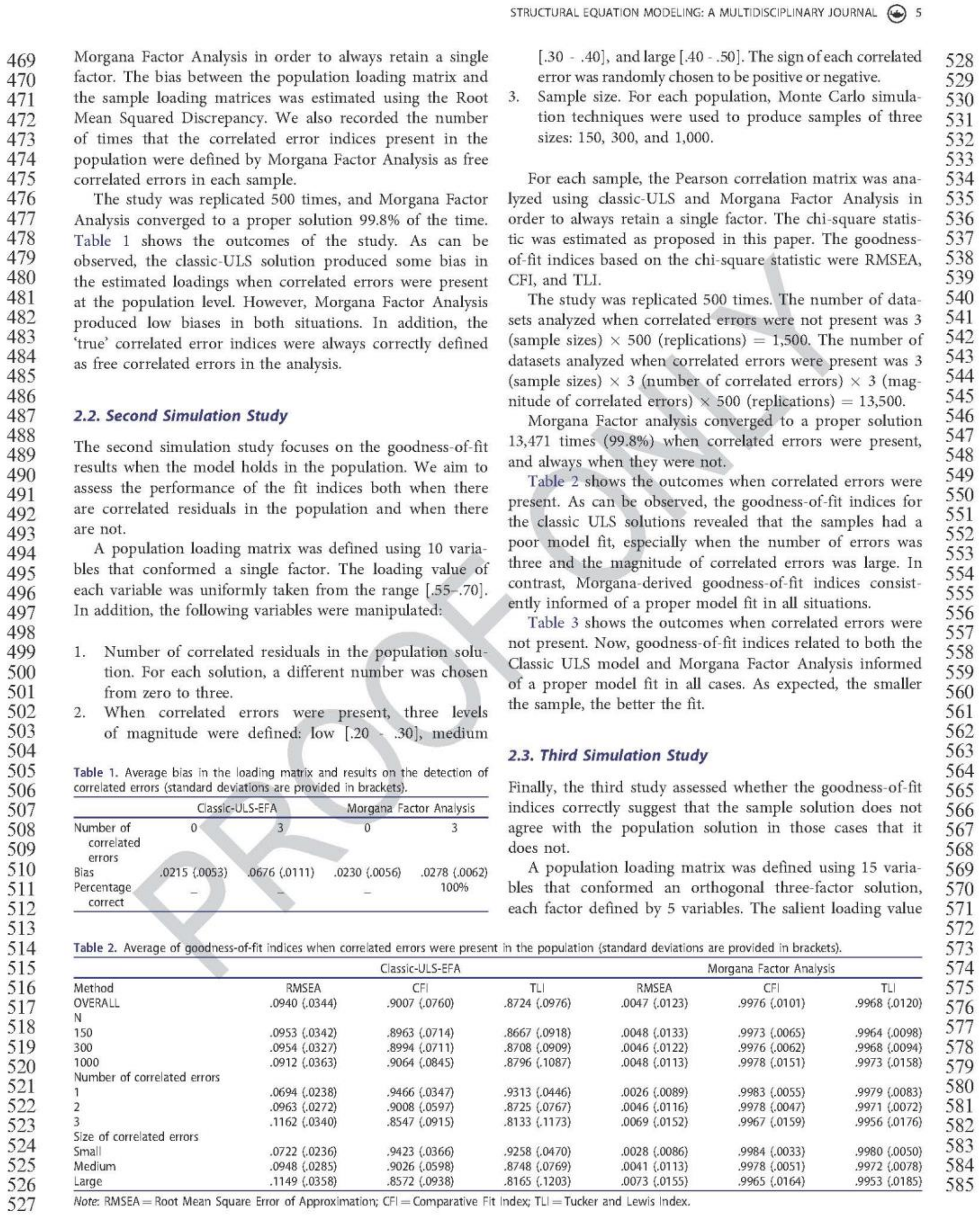
2. Simulation Studies

Three short simulation studies were planned to assess the performance of our approach. The first study assesses whether the loading estimates are correct and the 'true' correlated errors are properly detected. The second assesses whether the goodness-of-fit indices correctly indicate that the sample solution is compatible with the 'true' solution generated for the population. As the proposal is exploratory, this second study includes 'true' solutions with and without correlated errors. Finally, the third study assesses whether the goodness-of-fit indices correctly suggest that the sample solution is incompatible with the solution generated for the populationwhen the sample specification does not match it.

2.1. First Study

A population loading matrix was defined using 10 variables that conformed a single-factor solution. The loading of each variable was uniformly taken from the range [.55–.70]. Two types of solution were defined. In the first type, all correlated errors were defined to be zero. In the second, three correlated errors were uniformly and randomly chosen from the range [.40–.50]. The sign of each correlated error was randomly chosen to be positive or negative. Based on the population solution, Monte Carlo simulation techniques were used to generate samples with $N = 1{,}000$.

For each sample, the Pearson correlation matrix was analyzed using (a) the classic model (2) fitted by ULS, and (b)



Morgana Factor Analysis in order to always retain a single factor. The bias between the population loading matrix and the sample loading matrices was estimated using the Root Mean Squared Discrepancy. We also recorded the number of times that the correlated error indices present in the population were defined by Morgana Factor Analysis as free correlated errors in each sample.

The study was replicated 500 times, and Morgana Factor Analysis converged to a proper solution 99.8% of the time. Table 1 shows the outcomes of the study. As can be observed, the classic-ULS solution produced some bias in the estimated loadings when correlated errors were present at the population level. However, Morgana Factor Analysis produced low biases in both situations. In addition, the 'true' correlated error indices were always correctly defined as free correlated errors in the analysis.

2.2. Second Simulation Study

The second simulation study focuses on the goodness-of-fit results when the model holds in the population. We aim to assess the performance of the fit indices both when there are correlated residuals in the population and when there are not.

A population loading matrix was defined using 10 variables that conformed a single factor. The loading value of each variable was uniformly taken from the range [.55–.70]. In addition, the following variables were manipulated:

1. Number of correlated residuals in the population solution. For each solution, a different number was chosen from zero to three.
2. When correlated errors were present, three levels of magnitude were defined: low [.20 - .30], medium [.30 - .40], and large [.40 - .50]. The sign of each correlated error was randomly chosen to be positive or negative.
3. Sample size. For each population, Monte Carlo simulation techniques were used to produce samples of three sizes: 150, 300, and 1,000.

Table 1. Average bias in the loading matrix and results on the detection of correlated errors (standard deviations are provided in brackets).

	Classic-ULS-EFA		Morgana Factor Analysis	
Number of correlated errors	0	3	0	3
Bias	.0215 (.0053)	.0676 (.0111)	.0230 (.0056)	.0278 (.0062)
Percentage correct	–	–	–	100%

For each sample, the Pearson correlation matrix was analyzed using classic-ULS and Morgana Factor Analysis in order to always retain a single factor. The chi-square statistic was estimated as proposed in this paper. The goodness-of-fit indices based on the chi-square statistic were RMSEA, CFI, and TLI.

The study was replicated 500 times. The number of datasets analyzed when correlated errors were not present was 3 (sample sizes) × 500 (replications) = 1,500. The number of datasets analyzed when correlated errors were present was 3 (sample sizes) × 3 (number of correlated errors) × 3 (magnitude of correlated errors) × 500 (replications) = 13,500.

Morgana Factor analysis converged to a proper solution 13,471 times (99.8%) when correlated errors were present, and always when they were not.

Table 2 shows the outcomes when correlated errors were present. As can be observed, the goodness-of-fit indices for the classic ULS solutions revealed that the samples had a poor model fit, especially when the number of errors was three and the magnitude of correlated errors was large. In contrast, Morgana-derived goodness-of-fit indices consistently informed of a proper model fit in all situations.

Table 3 shows the outcomes when correlated errors were not present. Now, goodness-of-fit indices related to both the Classic ULS model and Morgana Factor Analysis informed of a proper model fit in all cases. As expected, the smaller the sample, the better the fit.

2.3. Third Simulation Study

Finally, the third study assessed whether the goodness-of-fit indices correctly suggest that the sample solution does not agree with the population solution in those cases that it does not.

A population loading matrix was defined using 15 variables that conformed an orthogonal three-factor solution, each factor defined by 5 variables. The salient loading value

Table 2. Average of goodness-of-fit indices when correlated errors were present in the population (standard deviations are provided in brackets).

	Classic-ULS-EFA			Morgana Factor Analysis		
Method	RMSEA	CFI	TLI	RMSEA	CFI	TLI
OVERALL	.0940 (.0344)	.9007 (.0760)	.8724 (.0976)	.0047 (.0123)	.9976 (.0101)	.9968 (.0120)
N						
150	.0953 (.0342)	.8963 (.0714)	.8667 (.0918)	.0048 (.0133)	.9973 (.0065)	.9964 (.0098)
300	.0954 (.0327)	.8994 (.0711)	.8708 (.0909)	.0046 (.0122)	.9976 (.0062)	.9968 (.0094)
1000	.0912 (.0363)	.9064 (.0845)	.8796 (.1087)	.0048 (.0113)	.9978 (.0151)	.9973 (.0158)
Number of correlated errors						
1	.0694 (.0238)	.9466 (.0347)	.9313 (.0446)	.0026 (.0089)	.9983 (.0055)	.9979 (.0083)
2	.0963 (.0272)	.9008 (.0597)	.8725 (.0767)	.0046 (.0116)	.9978 (.0047)	.9971 (.0072)
3	.1162 (.0340)	.8547 (.0915)	.8133 (.1173)	.0069 (.0152)	.9967 (.0159)	.9956 (.0176)
Size of correlated errors						
Small	.0722 (.0236)	.9423 (.0366)	.9258 (.0470)	.0028 (.0086)	.9984 (.0033)	.9980 (.0050)
Medium	.0948 (.0285)	.9026 (.0598)	.8748 (.0769)	.0041 (.0113)	.9978 (.0051)	.9972 (.0078)
Large	.1149 (.0358)	.8572 (.0938)	.8165 (.1203)	.0073 (.0155)	.9965 (.0164)	.9953 (.0185)

Note: RMSEA = Root Mean Square Error of Approximation; CFI = Comparative Fit Index; TLI = Tucker and Lewis Index.

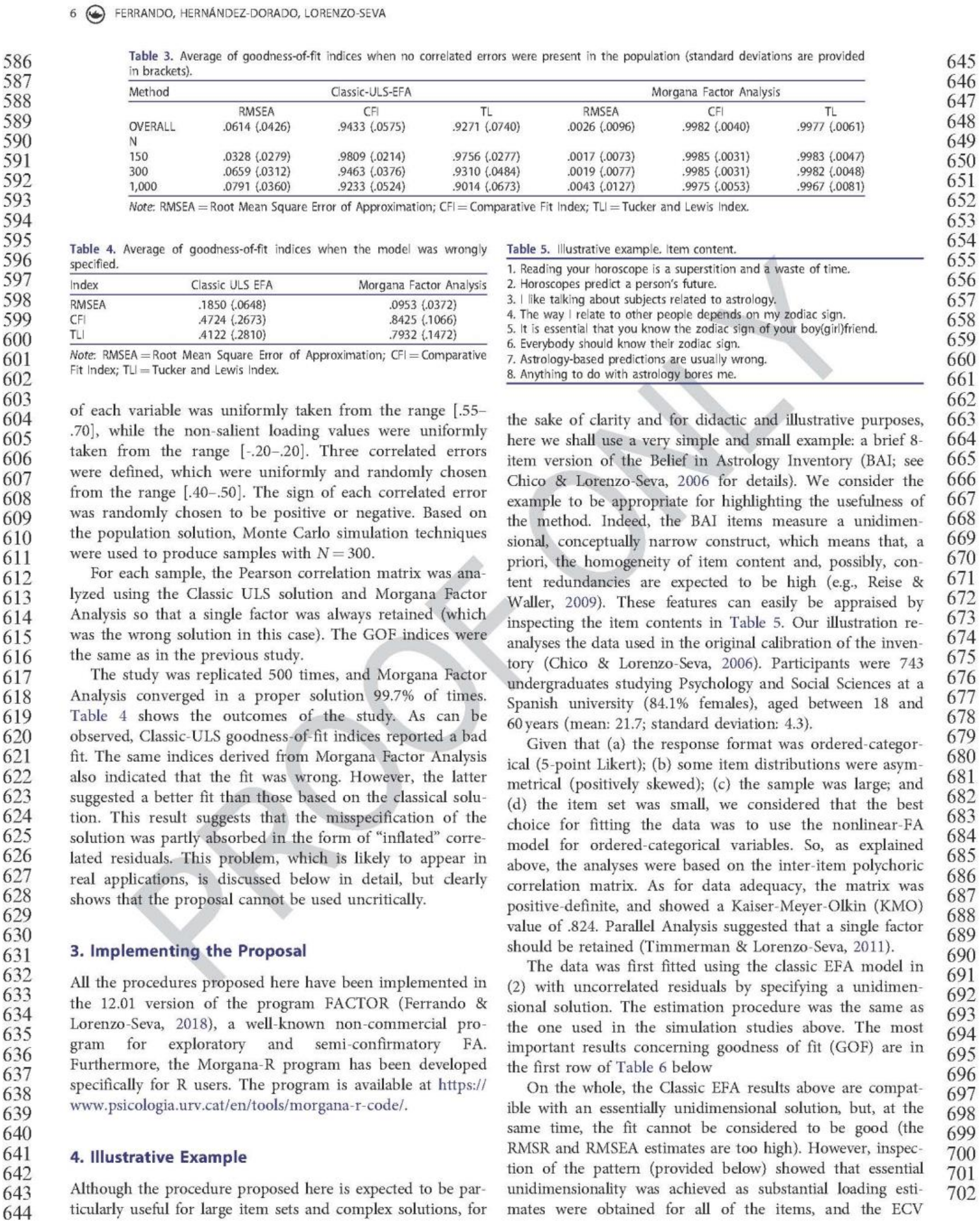


Table 3. Average of goodness-of-fit indices when no correlated errors were present in the population (standard deviations are provided in brackets).

Method	Classic-ULS-EFA			Morgana Factor Analysis		
	RMSEA	CFI	TL	RMSEA	CFI	TL
OVERALL	.0614 (.0426)	.9433 (.0575)	.9271 (.0740)	.0026 (.0096)	.9982 (.0040)	.9977 (.0061)
N						
150	.0328 (.0279)	.9809 (.0214)	.9756 (.0277)	.0017 (.0073)	.9985 (.0031)	.9983 (.0047)
300	.0659 (.0312)	.9463 (.0376)	.9310 (.0484)	.0019 (.0077)	.9985 (.0031)	.9982 (.0048)
1,000	.0791 (.0360)	.9233 (.0524)	.9014 (.0673)	.0043 (.0127)	.9975 (.0053)	.9967 (.0081)

Note: RMSEA = Root Mean Square Error of Approximation; CFI = Comparative Fit Index; TLI = Tucker and Lewis Index.

Table 4. Average of goodness-of-fit indices when the model was wrongly specified.

Index	Classic ULS EFA	Morgana Factor Analysis
RMSEA	.1850 (.0648)	.0953 (.0372)
CFI	.4724 (.2673)	.8425 (.1066)
TLI	.4122 (.2810)	.7932 (.1472)

Note: RMSEA = Root Mean Square Error of Approximation; CFI = Comparative Fit Index; TLI = Tucker and Lewis Index.

Table 5. Illustrative example. Item content.

1. Reading your horoscope is a superstition and a waste of time.
2. Horoscopes predict a person's future.
3. I like talking about subjects related to astrology.
4. The way I relate to other people depends on my zodiac sign.
5. It is essential that you know the zodiac sign of your boy(girl)friend.
6. Everybody should know their zodiac sign.
7. Astrology-based predictions are usually wrong.
8. Anything to do with astrology bores me.

of each variable was uniformly taken from the range [.55–.70], while the non-salient loading values were uniformly taken from the range [-.20–.20]. Three correlated errors were defined, which were uniformly and randomly chosen from the range [.40–.50]. The sign of each correlated error was randomly chosen to be positive or negative. Based on the population solution, Monte Carlo simulation techniques were used to produce samples with $N = 300$.

For each sample, the Pearson correlation matrix was analyzed using the Classic ULS solution and Morgana Factor Analysis so that a single factor was always retained (which was the wrong solution in this case). The GOF indices were the same as in the previous study.

The study was replicated 500 times, and Morgana Factor Analysis converged in a proper solution 99.7% of times. Table 4 shows the outcomes of the study. As can be observed, Classic-ULS goodness-of-fit indices reported a bad fit. The same indices derived from Morgana Factor Analysis also indicated that the fit was wrong. However, the latter suggested a better fit than those based on the classical solution. This result suggests that the misspecification of the solution was partly absorbed in the form of "inflated" correlated residuals. This problem, which is likely to appear in real applications, is discussed below in detail, but clearly shows that the proposal cannot be used uncritically.

3. Implementing the Proposal

All the procedures proposed here have been implemented in the 12.01 version of the program FACTOR (Ferrando & Lorenzo-Seva, 2018), a well-known non-commercial program for exploratory and semi-confirmatory FA. Furthermore, the Morgana-R program has been developed specifically for R users. The program is available at https://www.psicologia.urv.cat/en/tools/morgana-r-code/.

4. Illustrative Example

Although the procedure proposed here is expected to be particularly useful for large item sets and complex solutions, for the sake of clarity and for didactic and illustrative purposes, here we shall use a very simple and small example: a brief 8-item version of the Belief in Astrology Inventory (BAI; see Chico & Lorenzo-Seva, 2006 for details). We consider the example to be appropriate for highlighting the usefulness of the method. Indeed, the BAI items measure a unidimensional, conceptually narrow construct, which means that, a priori, the homogeneity of item content and, possibly, content redundancies are expected to be high (e.g., Reise & Waller, 2009). These features can easily be appraised by inspecting the item contents in Table 5. Our illustration reanalyses the data used in the original calibration of the inventory (Chico & Lorenzo-Seva, 2006). Participants were 743 undergraduates studying Psychology and Social Sciences at a Spanish university (84.1% females), aged between 18 and 60 years (mean: 21.7; standard deviation: 4.3).

Given that (a) the response format was ordered-categorical (5-point Likert); (b) some item distributions were asymmetrical (positively skewed); (c) the sample was large; and (d) the item set was small, we considered that the best choice for fitting the data was to use the nonlinear-FA model for ordered-categorical variables. So, as explained above, the analyses were based on the inter-item polychoric correlation matrix. As for data adequacy, the matrix was positive-definite, and showed a Kaiser-Meyer-Olkin (KMO) value of .824. Parallel Analysis suggested that a single factor should be retained (Timmerman & Lorenzo-Seva, 2011).

The data was first fitted using the classic EFA model in (2) with uncorrelated residuals by specifying a unidimensional solution. The estimation procedure was the same as the one used in the simulation studies above. The most important results concerning goodness of fit (GOF) are in the first row of Table 6 below

On the whole, the Classic EFA results above are compatible with an essentially unidimensional solution, but, at the same time, the fit cannot be considered to be good (the RMSR and RMSEA estimates are too high). However, inspection of the pattern (provided below) showed that essential unidimensionality was achieved as substantial loading estimates were obtained for all of the items, and the ECV

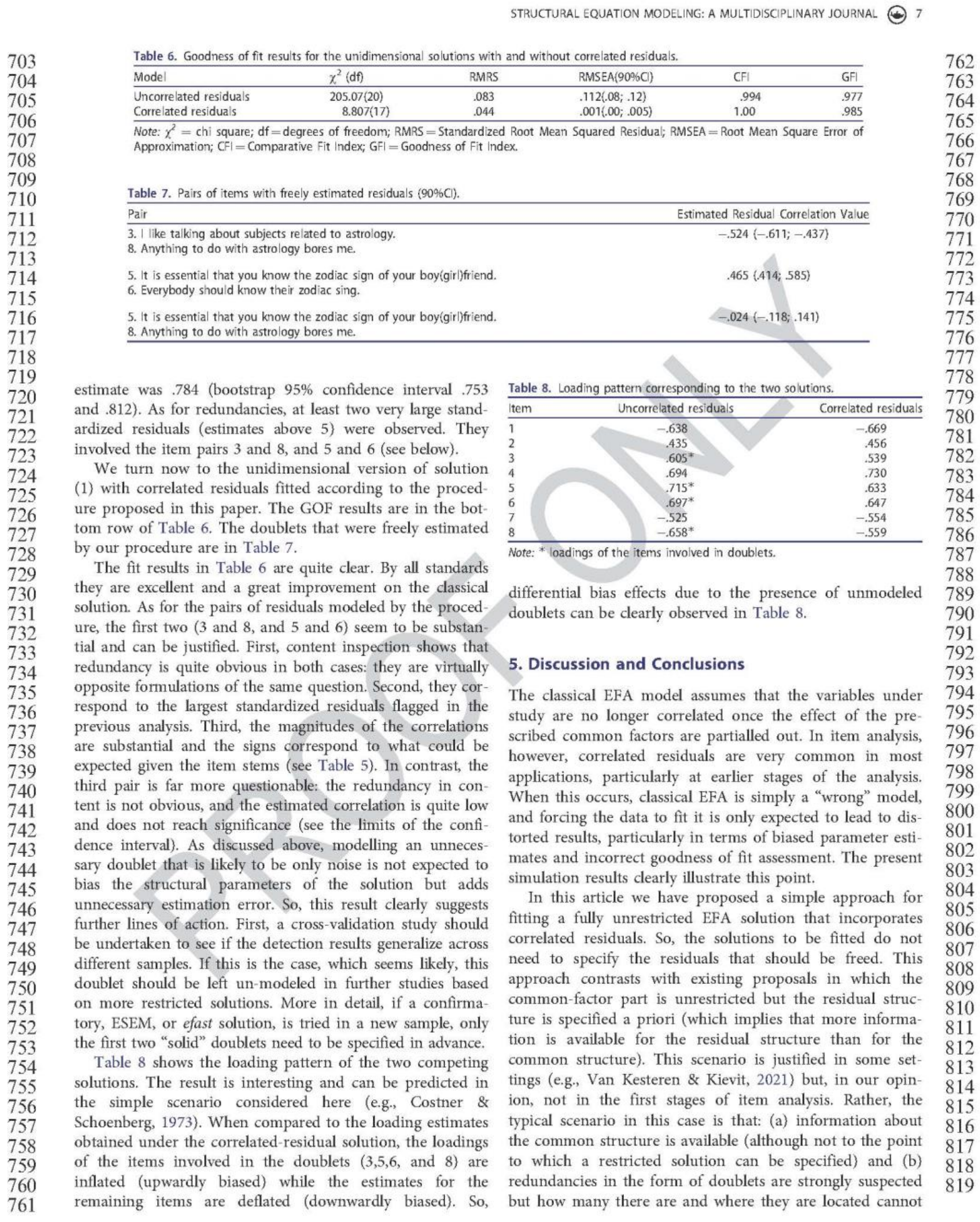


Table 6. Goodness of fit results for the unidimensional solutions with and without correlated residuals.

Model	χ^2 (df)	RMRS	RMSEA(90%CI)	CFI	GFI
Uncorrelated residuals	205.07(20)	.083	.112(.08; .12)	.994	.977
Correlated residuals	8.807(17)	.044	.001(.00; .005)	1.00	.985

Note: χ^2 = chi square; df = degrees of freedom; RMRS = Standardized Root Mean Squared Residual; RMSEA = Root Mean Square Error of Approximation; CFI = Comparative Fit Index; GFI = Goodness of Fit Index.

Table 7. Pairs of items with freely estimated residuals (90%CI).

Pair	Estimated Residual Correlation Value
3. I like talking about subjects related to astrology. 8. Anything to do with astrology bores me.	−.524 (−.611; −.437)
5. It is essential that you know the zodiac sign of your boy(girl)friend. 6. Everybody should know their zodiac sing.	.465 (.414; .585)
5. It is essential that you know the zodiac sign of your boy(girl)friend. 8. Anything to do with astrology bores me.	−.024 (−.118; .141)

estimate was .784 (bootstrap 95% confidence interval .753 and .812). As for redundancies, at least two very large standardized residuals (estimates above 5) were observed. They involved the item pairs 3 and 8, and 5 and 6 (see below).

We turn now to the unidimensional version of solution (1) with correlated residuals fitted according to the procedure proposed in this paper. The GOF results are in the bottom row of Table 6. The doublets that were freely estimated by our procedure are in Table 7.

The fit results in Table 6 are quite clear. By all standards they are excellent and a great improvement on the classical solution. As for the pairs of residuals modeled by the procedure, the first two (3 and 8, and 5 and 6) seem to be substantial and can be justified. First, content inspection shows that redundancy is quite obvious in both cases: they are virtually opposite formulations of the same question. Second, they correspond to the largest standardized residuals flagged in the previous analysis. Third, the magnitudes of the correlations are substantial and the signs correspond to what could be expected given the item stems (see Table 5). In contrast, the third pair is far more questionable: the redundancy in content is not obvious, and the estimated correlation is quite low and does not reach significance (see the limits of the confidence interval). As discussed above, modelling an unnecessary doublet that is likely to be only noise is not expected to bias the structural parameters of the solution but adds unnecessary estimation error. So, this result clearly suggests further lines of action. First, a cross-validation study should be undertaken to see if the detection results generalize across different samples. If this is the case, which seems likely, this doublet should be left un-modeled in further studies based on more restricted solutions. More in detail, if a confirmatory, ESEM, or *efast* solution, is tried in a new sample, only the first two "solid" doublets need to be specified in advance.

Table 8 shows the loading pattern of the two competing solutions. The result is interesting and can be predicted in the simple scenario considered here (e.g., Costner & Schoenberg, 1973). When compared to the loading estimates obtained under the correlated-residual solution, the loadings of the items involved in the doublets (3,5,6, and 8) are inflated (upwardly biased) while the estimates for the remaining items are deflated (downwardly biased). So, differential bias effects due to the presence of unmodeled doublets can be clearly observed in Table 8.

Table 8. Loading pattern corresponding to the two solutions.

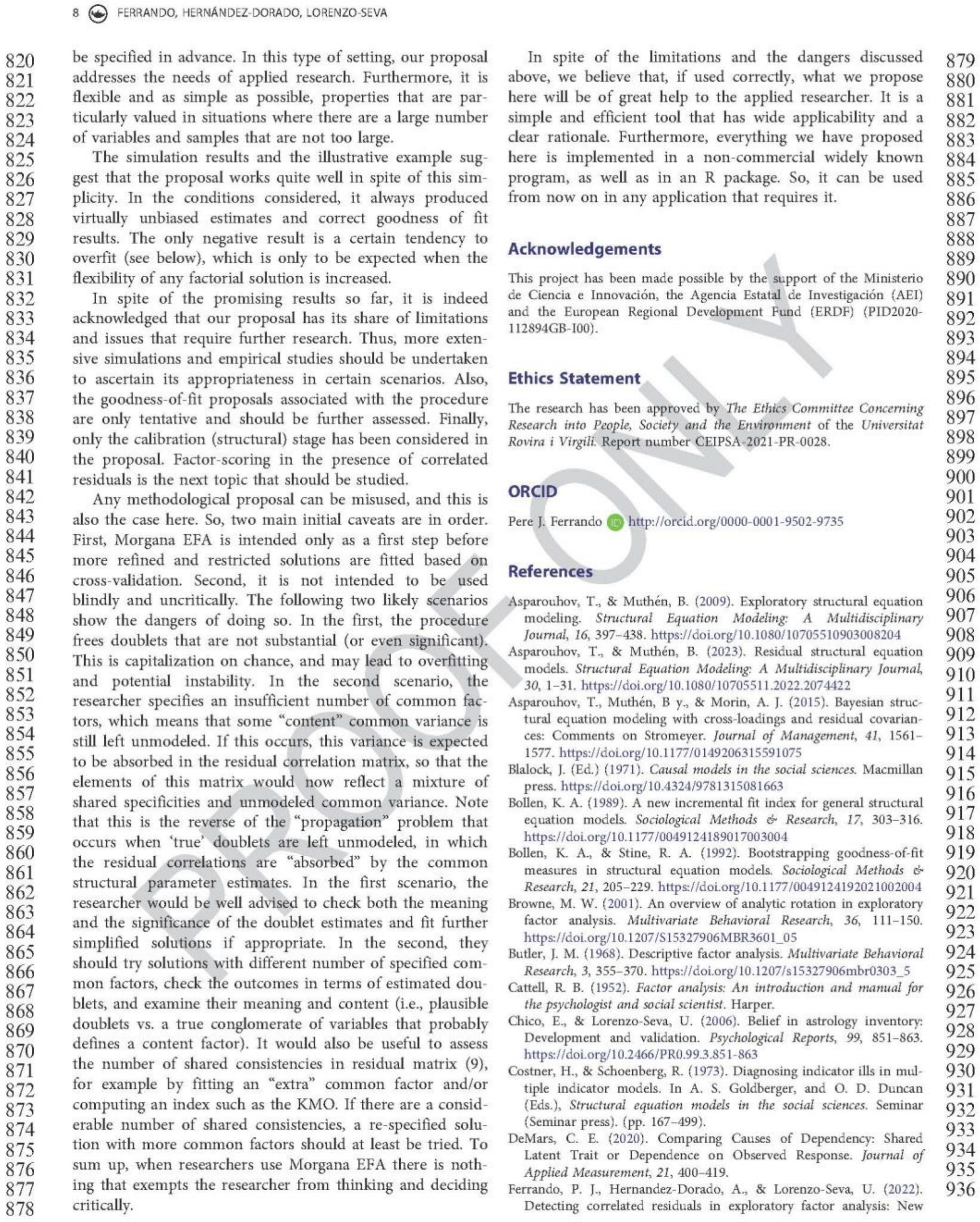
Item	Uncorrelated residuals	Correlated residuals
1	−.638	−.669
2	.435	.456
3	.605*	.539
4	.694	.730
5	.715*	.633
6	.697*	.647
7	−.525	−.554
8	−.658*	−.559

Note: * loadings of the items involved in doublets.

5. Discussion and Conclusions

The classical EFA model assumes that the variables under study are no longer correlated once the effect of the prescribed common factors are partialled out. In item analysis, however, correlated residuals are very common in most applications, particularly at earlier stages of the analysis. When this occurs, classical EFA is simply a "wrong" model, and forcing the data to fit it is only expected to lead to distorted results, particularly in terms of biased parameter estimates and incorrect goodness of fit assessment. The present simulation results clearly illustrate this point.

In this article we have proposed a simple approach for fitting a fully unrestricted EFA solution that incorporates correlated residuals. So, the solutions to be fitted do not need to specify the residuals that should be freed. This approach contrasts with existing proposals in which the common-factor part is unrestricted but the residual structure is specified a priori (which implies that more information is available for the residual structure than for the common structure). This scenario is justified in some settings (e.g., Van Kesteren & Kievit, 2021) but, in our opinion, not in the first stages of item analysis. Rather, the typical scenario in this case is that: (a) information about the common structure is available (although not to the point to which a restricted solution can be specified) and (b) redundancies in the form of doublets are strongly suspected but how many there are and where they are located cannot



be specified in advance. In this type of setting, our proposal addresses the needs of applied research. Furthermore, it is flexible and as simple as possible, properties that are particularly valued in situations where there are a large number of variables and samples that are not too large.

The simulation results and the illustrative example suggest that the proposal works quite well in spite of this simplicity. In the conditions considered, it always produced virtually unbiased estimates and correct goodness of fit results. The only negative result is a certain tendency to overfit (see below), which is only to be expected when the flexibility of any factorial solution is increased.

In spite of the promising results so far, it is indeed acknowledged that our proposal has its share of limitations and issues that require further research. Thus, more extensive simulations and empirical studies should be undertaken to ascertain its appropriateness in certain scenarios. Also, the goodness-of-fit proposals associated with the procedure are only tentative and should be further assessed. Finally, only the calibration (structural) stage has been considered in the proposal. Factor-scoring in the presence of correlated residuals is the next topic that should be studied.

Any methodological proposal can be misused, and this is also the case here. So, two main initial caveats are in order. First, Morgana EFA is intended only as a first step before more refined and restricted solutions are fitted based on cross-validation. Second, it is not intended to be used blindly and uncritically. The following two likely scenarios show the dangers of doing so. In the first, the procedure frees doublets that are not substantial (or even significant). This is capitalization on chance, and may lead to overfitting and potential instability. In the second scenario, the researcher specifies an insufficient number of common factors, which means that some "content" common variance is still left unmodeled. If this occurs, this variance is expected to be absorbed in the residual correlation matrix, so that the elements of this matrix would now reflect a mixture of shared specificities and unmodeled common variance. Note that this is the reverse of the "propagation" problem that occurs when 'true' doublets are left unmodeled, in which the residual correlations are "absorbed" by the common structural parameter estimates. In the first scenario, the researcher would be well advised to check both the meaning and the significance of the doublet estimates and fit further simplified solutions if appropriate. In the second, they should try solutions with different number of specified common factors, check the outcomes in terms of estimated doublets, and examine their meaning and content (i.e., plausible doublets vs. a true conglomerate of variables that probably defines a content factor). It would also be useful to assess the number of shared consistencies in residual matrix (9), for example by fitting an "extra" common factor and/or computing an index such as the KMO. If there are a considerable number of shared consistencies, a re-specified solution with more common factors should at least be tried. To sum up, when researchers use Morgana EFA there is nothing that exempts the researcher from thinking and deciding critically.

In spite of the limitations and the dangers discussed above, we believe that, if used correctly, what we propose here will be of great help to the applied researcher. It is a simple and efficient tool that has wide applicability and a clear rationale. Furthermore, everything we have proposed here is implemented in a non-commercial widely known program, as well as in an R package. So, it can be used from now on in any application that requires it.

Acknowledgements

This project has been made possible by the support of the Ministerio de Ciencia e Innovación, the Agencia Estatal de Investigación (AEI) and the European Regional Development Fund (ERDF) (PID2020-112894GB-I00).

Ethics Statement

The research has been approved by *The Ethics Committee Concerning Research into People, Society and the Environment* of the *Universitat Rovira i Virgili*. Report number CEIPSA-2021-PR-0028.

ORCID

Pere J. Ferrando http://orcid.org/0000-0001-9502-9735

References

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*, 397–438. https://doi.org/10.1080/10705510903008204

Asparouhov, T., & Muthén, B. (2023). Residual structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *30*, 1–31. https://doi.org/10.1080/10705511.2022.2074422

Asparouhov, T., Muthén, B y., & Morin, A. J. (2015). Bayesian structural equation modeling with cross-loadings and residual covariances: Comments on Stromeyer. *Journal of Management*, *41*, 1561–1577. https://doi.org/10.1177/0149206315591075

Blalock, J. (Ed.) (1971). *Causal models in the social sciences.* Macmillan press. https://doi.org/10.4324/9781315081663

Bollen, K. A. (1989). A new incremental fit index for general structural equation models. *Sociological Methods & Research*, *17*, 303–316. https://doi.org/10.1177/0049124189017003004

Bollen, K. A., & Stine, R. A. (1992). Bootstrapping goodness-of-fit measures in structural equation models. *Sociological Methods & Research*, *21*, 205–229. https://doi.org/10.1177/0049124192021002004

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*, *36*, 111–150. https://doi.org/10.1207/S15327906MBR3601_05

Butler, J. M. (1968). Descriptive factor analysis. *Multivariate Behavioral Research*, *3*, 355–370. https://doi.org/10.1207/s15327906mbr0303_5

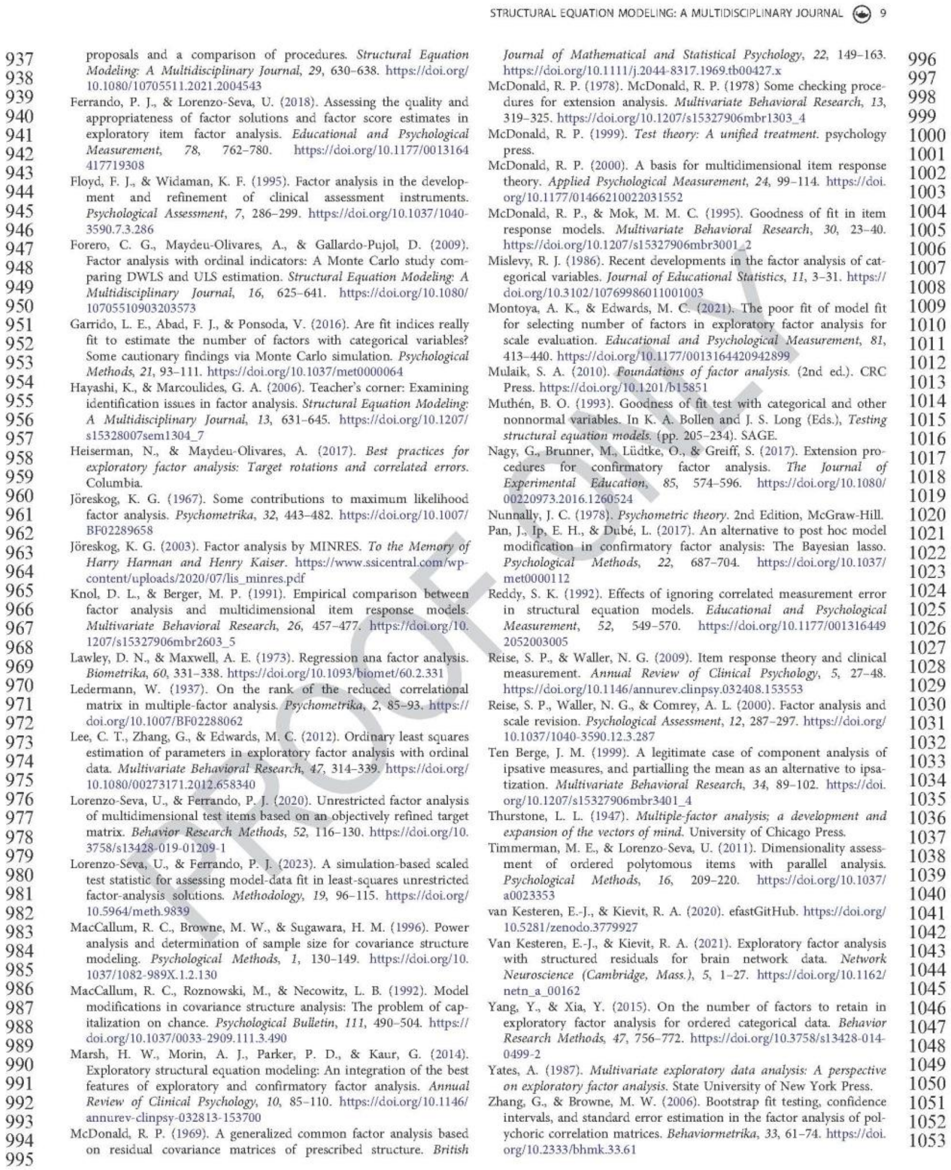
Cattell, R. B. (1952). *Factor analysis: An introduction and manual for the psychologist and social scientist*. Harper.

Chico, E., & Lorenzo-Seva, U. (2006). Belief in astrology inventory: Development and validation. *Psychological Reports*, *99*, 851–863. https://doi.org/10.2466/PR0.99.3.851-863

Costner, H., & Schoenberg, R. (1973). Diagnosing indicator ills in multiple indicator models. In A. S. Goldberger, and O. D. Duncan (Eds.), *Structural equation models in the social sciences*. Seminar (Seminar press). (pp. 167–499).

DeMars, C. E. (2020). Comparing Causes of Dependency: Shared Latent Trait or Dependence on Observed Response. *Journal of Applied Measurement*, *21*, 400–419.

Ferrando, P. J., Hernandez-Dorado, A., & Lorenzo-Seva, U. (2022). Detecting correlated residuals in exploratory factor analysis: New



proposals and a comparison of procedures. *Structural Equation Modeling: A Multidisciplinary Journal*, *29*, 630–638. https://doi.org/10.1080/10705511.2021.2004543

Ferrando, P. J., & Lorenzo-Seva, U. (2018). Assessing the quality and appropriateness of factor solutions and factor score estimates in exploratory item factor analysis. *Educational and Psychological Measurement*, *78*, 762–780. https://doi.org/10.1177/0013164417719308

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment*, *7*, 286–299. https://doi.org/10.1037/1040-3590.7.3.286

Forero, C. G., Maydeu-Olivares, A., & Gallardo-Pujol, D. (2009). Factor analysis with ordinal indicators: A Monte Carlo study comparing DWLS and ULS estimation. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*, 625–641. https://doi.org/10.1080/10705510903203573

Garrido, L. E., Abad, F. J., & Ponsoda, V. (2016). Are fit indices really fit to estimate the number of factors with categorical variables? Some cautionary findings via Monte Carlo simulation. *Psychological Methods*, *21*, 93–111. https://doi.org/10.1037/met0000064

Hayashi, K., & Marcoulides, G. A. (2006). Teacher's corner: Examining identification issues in factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *13*, 631–645. https://doi.org/10.1207/s15328007sem1304_7

Heiserman, N., & Maydeu-Olivares, A. (2017). *Best practices for exploratory factor analysis: Target rotations and correlated errors*. Columbia.

Jöreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika*, *32*, 443–482. https://doi.org/10.1007/BF02289658

Jöreskog, K. G. (2003). Factor analysis by MINRES. *To the Memory of Harry Harman and Henry Kaiser*. https://www.ssicentral.com/wp-content/uploads/2020/07/lis_minres.pdf

Knol, D. L., & Berger, M. P. (1991). Empirical comparison between factor analysis and multidimensional item response models. *Multivariate Behavioral Research*, *26*, 457–477. https://doi.org/10.1207/s15327906mbr2603_5

Lawley, D. N., & Maxwell, A. E. (1973). Regression ana factor analysis. *Biometrika*, *60*, 331–338. https://doi.org/10.1093/biomet/60.2.331

Ledermann, W. (1937). On the rank of the reduced correlational matrix in multiple-factor analysis. *Psychometrika*, *2*, 85–93. https://doi.org/10.1007/BF02288062

Lee, C. T., Zhang, G., & Edwards, M. C. (2012). Ordinary least squares estimation of parameters in exploratory factor analysis with ordinal data. *Multivariate Behavioral Research*, *47*, 314–339. https://doi.org/10.1080/00273171.2012.658340

Lorenzo-Seva, U., & Ferrando, P. J. (2020). Unrestricted factor analysis of multidimensional test items based on an objectively refined target matrix. *Behavior Research Methods*, *52*, 116–130. https://doi.org/10.3758/s13428-019-01209-1

Lorenzo-Seva, U., & Ferrando, P. J. (2023). A simulation-based scaled test statistic for assessing model-data fit in least-squares unrestricted factor-analysis solutions. *Methodology*, *19*, 96–115. https://doi.org/10.5964/meth.9839

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, *1*, 130–149. https://doi.org/10.1037/1082-989X.1.2.130

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin*, *111*, 490–504. https://doi.org/10.1037/0033-2909.111.3.490

Marsh, H. W., Morin, A. J., Parker, P. D., & Kaur, G. (2014). Exploratory structural equation modeling: An integration of the best features of exploratory and confirmatory factor analysis. *Annual Review of Clinical Psychology*, *10*, 85–110. https://doi.org/10.1146/annurev-clinpsy-032813-153700

McDonald, R. P. (1969). A generalized common factor analysis based on residual covariance matrices of prescribed structure. *British Journal of Mathematical and Statistical Psychology*, *22*, 149–163. https://doi.org/10.1111/j.2044-8317.1969.tb00427.x

McDonald, R. P. (1978). McDonald, R. P. (1978) Some checking procedures for extension analysis. *Multivariate Behavioral Research*, *13*, 319–325. https://doi.org/10.1207/s15327906mbr1303_4

McDonald, R. P. (1999). *Test theory: A unified treatment*. psychology press.

McDonald, R. P. (2000). A basis for multidimensional item response theory. *Applied Psychological Measurement*, *24*, 99–114. https://doi.org/10.1177/01466210022031552

McDonald, R. P., & Mok, M. M. C. (1995). Goodness of fit in item response models. *Multivariate Behavioral Research*, *30*, 23–40. https://doi.org/10.1207/s15327906mbr3001_2

Mislevy, R. J. (1986). Recent developments in the factor analysis of categorical variables. *Journal of Educational Statistics*, *11*, 3–31. https://doi.org/10.3102/10769986011001003

Montoya, A. K., & Edwards, M. C. (2021). The poor fit of model fit for selecting number of factors in exploratory factor analysis for scale evaluation. *Educational and Psychological Measurement*, *81*, 413–440. https://doi.org/10.1177/0013164420942899

Mulaik, S. A. (2010). *Foundations of factor analysis*. (2nd ed.). CRC Press. https://doi.org/10.1201/b15851

Muthén, B. O. (1993). Goodness of fit test with categorical and other nonnormal variables. In K. A. Bollen and J. S. Long (Eds.), *Testing structural equation models*. (pp. 205–234). SAGE.

Nagy, G., Brunner, M., Lüdtke, O., & Greiff, S. (2017). Extension procedures for confirmatory factor analysis. *The Journal of Experimental Education*, *85*, 574–596. https://doi.org/10.1080/00220973.2016.1260524

Nunnally, J. C. (1978). *Psychometric theory*. 2nd Edition, McGraw-Hill.

Pan, J., Ip, E. H., & Dubé, L. (2017). An alternative to post hoc model modification in confirmatory factor analysis: The Bayesian lasso. *Psychological Methods*, *22*, 687–704. https://doi.org/10.1037/met0000112

Reddy, S. K. (1992). Effects of ignoring correlated measurement error in structural equation models. *Educational and Psychological Measurement*, *52*, 549–570. https://doi.org/10.1177/0013164492052003005

Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology*, *5*, 27–48. https://doi.org/10.1146/annurev.clinpsy.032408.153553

Reise, S. P., Waller, N. G., & Comrey, A. L. (2000). Factor analysis and scale revision. *Psychological Assessment*, *12*, 287–297. https://doi.org/10.1037/1040-3590.12.3.287

Ten Berge, J. M. (1999). A legitimate case of component analysis of ipsative measures, and partialling the mean as an alternative to ipsatization. *Multivariate Behavioral Research*, *34*, 89–102. https://doi.org/10.1207/s15327906mbr3401_4

Thurstone, L. L. (1947). *Multiple-factor analysis; a development and expansion of the vectors of mind*. University of Chicago Press.

Timmerman, M. E., & Lorenzo-Seva, U. (2011). Dimensionality assessment of ordered polytomous items with parallel analysis. *Psychological Methods*, *16*, 209–220. https://doi.org/10.1037/a0023353

van Kesteren, E.-J., & Kievit, R. A. (2020). efastGitHub. https://doi.org/10.5281/zenodo.3779927

Van Kesteren, E.-J., & Kievit, R. A. (2021). Exploratory factor analysis with structured residuals for brain network data. *Network Neuroscience (Cambridge, Mass.)*, *5*, 1–27. https://doi.org/10.1162/netn_a_00162

Yang, Y., & Xia, Y. (2015). On the number of factors to retain in exploratory factor analysis for ordered categorical data. *Behavior Research Methods*, *47*, 756–772. https://doi.org/10.3758/s13428-014-0499-2

Yates, A. (1987). *Multivariate exploratory data analysis: A perspective on exploratory factor analysis*. State University of New York Press.

Zhang, G., & Browne, M. W. (2006). Bootstrap fit testing, confidence intervals, and standard error estimation in the factor analysis of polychoric correlation matrices. *Behaviormetrika*, *33*, 61–74. https://doi.org/10.2333/bhmk.33.61

It's not so bad! The impact and consequences of correcting for acquiescence when correlated residuals are present

Hernández-Dorado, A; Ferrando, P.J. & Vigil-Colet, A.

In spite of the interest generated by controlling variance unrelated to the content in personality measures, few studies have assessed the combined impacts of two sources of error. **Method**: This article compares three control procedures in databases that simultaneously exhibit acquiescence (ACQ) and correlated residuals: the SIREN method (Navarro-Gonzalez, et al., in press; ACQ control), the MORGANA method (Ferrando et al., 2022; in press; correlated residuals control), and a combined double control method. A "control" procedure was also examined in which the presence of both determinants was ignored. **Results**: The findings reveal significant differences among the three control methods, with the ACQ control method and the combined method proving more effective. Moreover, when the residual variance shifts to the factor loadings, it appears to be captured by the ACQ correction method. **Conclusions**: The findings suggest recommending a mixed correction procedure for acquiescence and correlated residuals

A pesar del interés que suscita el control de la varianza no relacionada con el contenido en medidas de personalidad, pocos estudios han evaluado los impactos combinados de dos fuentes de error. **Método:** Este artículo compara tres procedimientos de control en bases de datos que presentan simultáneamente aquiescencia (ACQ) y residuales correlacionados: método SIREN (Navarro-Gonzalez, et al., in press; control de ACQ), método MORGANA (Ferrando et al., 2022; in press; control de residuales correlados) y un método doble de control combinado. También se examinó un procedimiento "control" en el que se ignoraba la presencia los dos determinantes. **Resultados:** Los resultados muestran diferencias significativas entre los tres métodos de control, siendo más eficientes el método de control de ACQ y el combinado. Además, cuando la varianza residual de desplaza a las cargas factoriales, esta parece ser captada por el método de corrección de ACQ. **Conclusiones:** Los hallazgos permiten recomendar un procedimiento de corrección mixto de aquiescencia y residuales correlacionados.

The measurement of personality variables via questionnaires presents a series of problems that have been extensively debated in the literature since the 30's. The general issue is that the response to a personality item can be potentially impacted by a series of determinants other than the 'content' intended to be measured (Cronbach, 1946). From a factor analytic (FA) view, when this is the case, two general problems can be expected to occur. First, a clear structure with the most realistic number of substantive factors cannot be attained. Second, individual factor score estimates derived from the structural solution cannot be univocally interpreted (Messick, 1995; Kline, 2015)

While there is an enormous amount of literature about the unwanted, non-content response determinants to personality items, studies focused on the joint impact of different sources are much rarer (e.g. Ferrando & Anguiano-Carrasco 2010). This state of affairs is somewhat surprising if we consider that, when responding to a personality item, several types of unwanted determinants are probably jointly operating, and that the overall impact can be the result of (perhaps) complex relations among them. Furthermore, the existing research on the topic and derived procedures (e.g. Ferrando et al, (2003) have considered variables from the same general type (mainly the joint occurrence of several response styles: acquiescence, social desirability or extreme response).

The present paper also focuses on the joint impact of two non-content response determinants. However, in contrast to the existing research above, these determinants are now from two different types: Response biases on the one hand and Method effects on the other. More specifically, we shall focus on acquiescence as the determinant in the first group, and local dependence (redundancy, correlated residuals) in the second.

Acquiescence

Acquiescence (ACQ) is one of the most studied response biases, either as the only distorting determinant (Bentler et al., 1971) or accompanied by others, mainly social desirability (SD) (Ferrando et al., 2009; Navarro-Gonzalez et al., 2016). It is generally viewed as a respondent-dependent response style, and, with regards to its impact, when operating tends to increase the correlations between items that are worded in the same direction but are not conceptually related (Podsakoff, 2003). Existing evidence suggests that about 6% of participants respond in a pronounced acquiescent or disacquiescent (DACQ) way (Hinz et al., 2007), and that about 4% of the variance of personality items is due to ACQ (Danner et al., 2015). In both cases, the percentages are far from being trivial, and therefore, the effects, detection and correction of ACQ have been extensively researched and documented (Baumgartner & Steenkamp, 2001; Billiet & McClendon, 2000; Savalei & Falk, 2014; Primi et al., 2019; de la Fuente & Abad, 2020). In psychometric terms, ACQ may affect the structural estimates at the calibration stage, distort the individual score estimates at the scoring stage, and bias the model-data fit assessment results. Focusing more specifically on calibration, the impact of ACQ can vary intra-individually depending on the measures or trait being measured (Ray, 1983); can cause biases in the invariance of the loadings on the content factor in comparative studies; and generate substantial bias in the estimation of stability coefficients and cross-lagged effects between variables over time in models based on panel data. (Billiet & Davidov, 2008).

Table 1.

Summary of some ACQ -control methods

Procedure	Author	Description	Pros	Cons
Ipsative method	Chan & Bentler (1993)	Developed for fitting factor models to data that have an ipsative structure,	Only feasible alternative if one wants to factor analyze measures with this structure	Weak functioning if there is a violation of the assumption of fully balanced scales and homogeneous ACQ loadings
General factor style Model	Billiet & McClendon (2000)	They developed a model that included a style factor that affected all items.	Easy to implement and improvement in goodness-of-fit indices compared to the uncorrected ACQ model.	The scale is required to be balanced and loadings on the style factor are required to be tau-equivalent
Unrestricted FA with Target Rotation	Ferrando et al., 2003	Unrestricted FA model in which ACQ is explicitly modeled as a secondary factor orthogonal to content.	Allows for differential ACQ loadings to be estimated.	Fully balanced scales are required.
RIFA	Maydeu-Olivares & Coffman, 2006	CFA model in which the additional ACQ factor is restricted to have equal loadings.	Easy to implement.	Tau-equivalence in ACQ factor loadings
Partially Balanced EFA	Lorenzo-Seva & Ferrando, 2009	Correction method derived from an adaptation of the rotation method (Lorenzo-Seva & Rodriguez-Fornells, 2006) that allows for the removal of variance due to ACQ in partially balanced scales.	Works also with partially balanced scales. Robust.	A minimal number of reversed items is required.
RI-EFA model	Aichholzer (2014)	A hybrid model that combines an EFA part where item-factor loadings are freely estimated and a restricted CFA part where item-factor loadings on the RI/ARS factor α are restricted to follow a predefined pattern	Can be extended to testing measurement invariance over subgroups or over time as well as to testing covariates of the RI/ARS factor and, hence, causes of such bias	Implementation is complex
Hybrid CFA-EFA (Siren)	Navarro-González et al. (in press)	Multi-stage procedure designed for fitting restricted FA solution in data matrices that have been cleaned from ACQ bias.	Allows restricted solutions to be fitted with the standard linear FA model or the non-linear graded-response model.	Sequential and conditional ad lib procedure that necessarily entails a loss of efficiency.

The results above justify the interest in measuring or controlling ACQ, and, so far, most of the existing procedures for these purposes fall within two broad categories: "A priori" methods linked to the design of the items, mainly the use of balanced scales (Ray, 1979). An "Ex post facto" methods mainly based on statistical control of the data. Table 1 summarizes de main control methods that have been developed in recent decades.

Several studies exist in which the pros and cons of the procedures in table 1 are compared in terms of performance (e.g. Savalei & Falk, 2014; Primi, et al., 2019; de la Fuente & Abad 2020). In general, the RIFA method usually emerges as the winner. However, this first place is more because of easiness of implementation than to real differences in effectiveness with the rest of the methods.

We briefly revise finally more complex studies in which the joint impact of ACQ and another response style has been assessed. So far, two "secondary" response styles have been considered: Extreme Response (ER) (Cheung & Rensvold, 2000; Weijters, et al., 2010; Park & Wu, 2019) and social desirability (SD) (e.g Hand & Brazzell, 1965; Ross & Mirowky, 1984; Ferrando & Anguiano-Carrasco, 2010). On the practical side, a procedure was proposed to control the bias caused by SD and ACQ simultaneously (Ferrando et al., 2009), which has been used in the construction of certain personality scales as OPERAS (Vigil-Colet et al., 2013) or INCA (Morales-Vives et al., 2019).

Local Dependence-Correlated Residuals

The terms "Local dependencies", "Correlated residuals", "Doublets" or "Shared specificities" refer to a common phenomenon which, in the present context, can be defined as follows. First, a pair (or a small group) of items continues to be related after the influence of the common content they measure has been partialed out. Second, this residual relation is due to other causes than additional common contents, such as context effects, redundancies in the evoked situation, wording similarities (Ferrando et al., 2022; Ferrando et al., in press). A main point here is that the causes just described are not linked to individual response tendencies or styles (as in ACQ) but to specific properties of the items. So, their existence is mostly related to the design of the measurement instrument.

While ACQ has been studied along multiple lines, research on residual correlations has focused above all on the convenience or not of allowing residuals to be modeled in factorial solutions. The effects of not controlling them, however, even when known, have been far less assessed (perhaps because researchers view this as a problem of test construction). As a result, residual analysis is rarely undertaken so far. With regards to these effects, they have been mostly assessed at the calibration stage, and are: (a) biased parameter estimates, and (b) distorted model-data fit assessment (Montoya & Edwards, 2021).

The detection of correlated residuals or doublets is mainly based on the inspection of the residual covariance matrix (Ferrando et al., 2022). In the case of a traditional exploratory factor analysis, where the residual covariances are forced to be zero, an un-modeled doublet can result on an increase of the corresponding fitted residual, an overall increase in the residual covariances or a propagation "shift" leading to an overestimation of the factor loadings involved in the doublet. This propagation effect can well make the doublet undetectable, so fitted residual inspection, despite being the fastest and simplest approach, is not the always the most appropriate. The partial-correlations method or the MORGANA method (Ferrando et al., 2022), despite being more complex, are expected to attain better results. More specifically, MORGANA, is derived from the concept of Expected Parameter Change (EPC; Saris, et al., 1987) and is able to minimize the propagation effects of substantial doublets to other residuals or to the factor loadings. MORGANA would include two indices: EREC and ENIDE. The first quantifies the amount of misspecification in the residual correlation; whereas, ENIDE is an auxiliary index that quantifies the extent to which the misspecification spreads and biases the loadings.

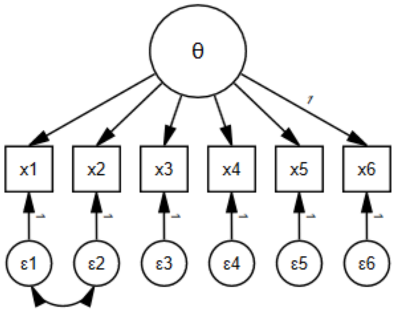
Once detected, the user must decide what to do with these residuals, and there are two obvious options: (a) remove one of the redundant items or (b) include the residuals in the model. Introducing correlated residuals in a model would increase the number of additional parameters leading to better model-data fit results. However, this improvement of fit is likely to imply capitalization of chance and a loss of replicability or reproducibility of the research results.

In summary, the research to date on unintended response determinants to personality items has gaps that need to be filled, and one of them is to determine the possible impact of two unwanted elements of different origin: ACQ on the one hand and the presence of correlated residuals on the other. The presence of each one of them separately is known to cause distortions in the calibration of the model, the estimation of the individual scores and the fit indices. So, it seems relevant to ask what effects can be caused by their joint occurrence.

The current goal of this study, therefore, is to examine the combined impact of ACQ and correlated residuals on the structural and score estimates of personality measures. And, in order to derive general predictions, the usual factor-analytic framework with residual correlations restricted to be zero will be used.

Figure 1.

Unidimensional model



From here on, the structure of this article is as follows. First, we shall derive some basic statistical predictions to provide the groundwork for our next steps. Next, a simulation study in which independent variables that are known to impact the structure and the goodness of fit of the model will be manipulated will be undertaken. The simulation will be completed with an empirical study based on personality data. Finally, we shall discuss the implications of the results found.

Basic Predictions

Consider a test that measures a personality trait (θ) and that is made up of n continuous response items. All of them present a response scale oriented in the same direction (e.g. 0: strongly disagree vs. 5: strongly agree), but half of the items are positively oriented and the other half are reverted. First, suppose that the scale is ACQ-free, but note that the residuals for items x1 and x2 are correlated (Figure 1). The base model is:

$$x_{ij} = \alpha_j\theta_i + \varepsilon_{ij} \qquad (1)$$

where x_{ij} is the standardized score of person *i* on item j, α is the factor loading, and ε is the measurement error. The factor (θ) is the trait that is scaled in a z-score metric (mean 0 and variance 1). Having determined the coefficients α of the factor θ, it is necessary to obtain the residual correlations. For items *j* and *k*, the residual covariance of those two items would be:

$$res_{jk} = \sigma_{jk} - \alpha_j\alpha_k \qquad (2)$$

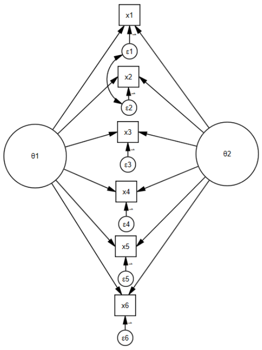
and, under the assumption that the residuals are uncorrelated, the observed correlation between a pair of items is given by (Harman, 1962; pp 120-1):

$$\sigma_{jk} = \alpha_j\alpha_k + res_{jk} = \alpha_j\alpha_k + 0 = \alpha_j\alpha_k \qquad (3)$$

where σ_{jk} is the covariance between items *j* and *k* and the αs are the loadings of items *j* and *k* of the content factor. However, if correlated residuals exist, but have been forced to be zero, the resulting omitted covariances will tend to be re-assigned and may (a) spread throughout the elements of the residual matrix or (b) bias the loading estimates of the content factor, specifically this second effect tend to focus on one of the two items involved in the doublet.

Figure 2.

The bidimensional for ACQ bias



Note. The bidimensional for ACQ bias (Ferrando & Lorenzo-Seva, 2010; Ferrando et al., 2003) that includes a residual correlation between item x1 and item x2. θ_1 represent the content factor, θ_2 represent the acquiescence factor and residual terms are shown as ε.

Taking now a step forward, let us consider a new model expanded from expression (1). It is an unrestricted FA model that was already proposed by Ferrando et al. (2003) and that includes two uncorrelated common factors: (a) a factor of content (θ_1) and (b) an acquiescence factor (θ_2) (Figure 2).

$$x_{ij} = \alpha_{j1}\theta_{i1} + \alpha_{j2}\theta_{i2} + \varepsilon_{ij} \tag{4}$$

The loadings on the content factor (α_{j1}) are considered balanced (half of the items will be positive and the other half negative) and larger than the loadings on the ACQ, and the loadings of the second factor are positive. Therefore, the sum of the content loadings will be null ($\sum_j^n \alpha_{j1} = 0$). The sum of the standardized scores becomes:

$$X_{ij} = \sum_j^n x_{ij} = \theta_{i2}\sum_j^n \alpha_{i2} + \sum_j^n \varepsilon_{ij} \tag{5}$$

and residual covariance according to (5), after fitting the model in a balanced scale is

$$res_{\,jk} = \alpha_{j2}\alpha_{k2} \tag{6}$$

that is, fitting a one-dimensional model like (1) to the inter-item covariance matrix is expected to identify the ACQ factor. Furthermore, if, (a) the scale is well balanced, and (b) as assumed the residuals correlations are all zero, then the loadings in this factor will be unbiased estimates of the item proneness to elicit ACQ. (Ferrando & Lorenzo-Seva 2009). However, if correlated residuals exist, they are expected to be absorbed in the estimated ACQ loadings that will then become biased. Overall, if the bidimensional model (4) with a content factor and an ACQ factor is fitted under balanced conditions but with non-zero correlated residuals, it could be assumed that, in order to keep $res_{\,jk}$ as close to zero as possible, the ACQ factor would absorb both the true ACQ and (to certain extent) any possible doublet that exist.

In light of these predictions, the aims of the present research are: (1) assessing the impact of correcting acquiescence in the estimated correlated residuals, and (2) observing how the use or omission of (a) ACQ estimation and (b) residual correction methods affects the estimation of content factor loadings and model-data fit results. This information will enable us to propose an informed procedural guide for scenarios containing a combined presence of variance unrelated to content.

Study 1: Monte Carlo Simulation Study 1

The first simulation study aimed to verify predictions made previously and test the effect of controlling and "eliminating" the variance due to acquiescence in the detection of correlated residuals. To do this, all samples were simulated under the two-dimensional model with one and/or two content factors (F1 and F2) and an acquiescence factor (ACQ) (Figure. 2). Depending on the number of factors, the number of items varied, with 6 items in the case of single-content factor models and 8 items in two-content factor models. In all cases, the factor loadings were completely balanced.

The study design was a full factorial 2 x 2 x 2 x 2 x 2, and the following variables were manipulated: (1) type of analysis: either sole residual control or combined ACQ and residual control; (2) the magnitude of acquiescence factor loadings (ACQ-L = .15 and .30); (3) the homogeneity of the ACQ loadings (ACQ-H); (4) location of items that correlate residually (within the same factor or in different factors); and (5) the sign of the items that residually correlate (items with the same sign or opposite sign). Out of the 32 databases with different characteristics that were simulated, half of them underwent a unique procedure for controlling correlated residuals using the MORGANA method. The other 16 databases were preprocessed by using the SIREN procedure (Navarro-Gonzalez et al., in press), firstly, variance due to ACQ was detected and eliminated, and subsequently, controlled correlated residuals were estimated from a bias-free correlation matrix. Both used methods are currently implemented in R (MORGANA is also available starting from FACTOR version 12.03.02).

The dependent variables were: (a) the number of doublets detected (either correctly or incorrectly) across the 500 replications, and (b) the estimated value of the EREC index for those items that truly formed doublets. From this information and, in order to examine the absorption effect, several contingency tables and ANOVAs were conducted to determine whether the number of doublets detected and the EREC index values depended or not on the prior correction of acquiescence. As a clarification regarding the number of detected doublets, EREC has proven to be a highly sensitive index, so only values greater than .20 were considered (Ferrando et al., 2022; in press).

The results of the first simulation study are summarized in Tables 2 and 3. Table 2 is a contingency table that assesses the number of doublets detected by the EREC index under different conditions. It can be observed that when the items that form the doublet are in the same location, absorption occurs in 100% of cases, regardless of whether the items have the same sign or opposite signs. However, the probability of absorption decreases by approximately 25% when the items are in different factors (recall that the simulation considers two orthogonal factors).

When there is no prior correction of ACQ, the results vary depending on whether the items have the same sign or opposite signs. When the two items that make up the doublet are positive, there is a tendency to overestimate the magnitude of the correlated residuals. It is interesting to note that, in this group, in only 0.73% of cases does the EREC index not detect the simulated doublet. This fact is especially noteworthy when compared to the condition in which the doublet falls on items with opposite signs, in which the number of false negatives reaches 23.5%. In the no-correction condition (of ACQ) in uncorrelated two-factor models, overestimation occurs in 100% of the cases. This contrasts significantly with the no-correction condition of ACQ in one-factor models, where overestimation is slightly lower and affects less than 65% of the sample (when the doublet involves items of the same sign and when it involves items of opposite signs).

The detection of doublets by EREC index in one-factor models that have been not corrected for ACQ is usually accurate; however, the analysis of variance revealed that the sign of the items had a significant impact, with a large effect size ($F_{(1)} = 8{,}241.1; p < .05 ; \eta^2 = .352$): in the case of doublets involving items with opposite signs, the number of false positives increased significantly. Nevertheless, in the case of two-factor models, false positives occurred in practically all replications.

Table 2.

Contingency table. Number of doublets detected by the EREC index

	1 factor				2 factors			
Nº D	Positive		Negative		Positive		Negative	
	C-ACQ	W-ACQ	C-ACQ	W-ACQ	C-ACQ	W-ACQ	C-ACQ	W-ACQ
0	4000 (100%)	29 (.73%)	4000 (100%)	941 (23.52%)	2942 (73.55%)		2973 (74.32%)	
1		1434 (35.85%)		726 (18.15%)	365 (9.13%)		329 (8.23%	
2		2537 (63.42%)		2333 (58.33%)	244 (6.1%)		253 (6.33%)	
3					449 (11.22%)	4000 (100%)	445 (11.12%)	4000 (100%)

Note. NºD = Number of Doublets; C-ACQ = Acquiescence control; W-ACQ=without prior acquiescence correction

Table 3 shows the ANOVA results concerned with the number of detected doublets. Only significant variables are shown. The results support the information provided in Table 2, indicating that the primary factor influencing the change in correlated residuals detection is the prior correction of acquiescence, with a large effect size ($\eta^2 = .58$).

Tabla 3.

Summary of ANOVA

	F	Gl	p- value	Effect size
Factor	27,190.0	1	< .05	.15*
Sign	132.2	1	< .05	<.001
ACQ-L	1,158.3	1	< .05	<.001
Correction	100,500.0	1	< .05	.58*
Factor:Sig	127.7	1	< .05	<.001
ACQ-L:Sign	106.6	1	< .05	<.001
Factor:Correction	6,095.0	1	< .05	.003
ACQ-L:Correction	1,647.0	1	< .05	<.001
Signo:Correction	127.8	1	< .05	<.001
Factor:acqh:acqsize	4.0	1	< .05	<.001
Factor: ACQ-L:Sign	141.7	1	< .05	<.001
Acqh: ACQ-L:Sign	7.5	1	< .05	<.001
Factor:ACQ-L:Correction	1,589.0	1	< .05	<.001
Azqh:ACQ-L:Correction	4.1	1	< .05	<.001
ACQ-L:Sign:Correction	141.8	1	< .05	<.001
Factor:ACQ-L:Sign:Correction	106.2	1	< .05	<.001
Factor:ACQ-H:ACQ-L:Sign:Correction	7.5	1	< .05	<.001

Note: * = large effect size

When data conforms a two-factor model, 100% of the detected residuals were false positives, whereas in the case of uncorrected one-factor models (remember that the corrected residuals were entirely absorbed), the analysis of variance detected that the sign of the items (items with the same sign or different signs) showed significant differences and a large effect size. ($F_{(1)} = 8{,}241.1$; $p < .05$; $\eta^2 = .352$).

Study 2: Montecarlo Simulation Study

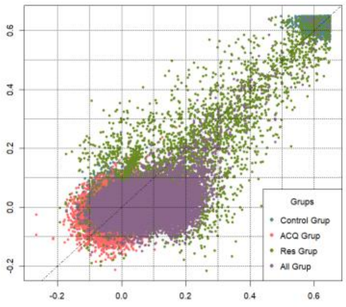
Once the absorption effect in correlated residuals when controlling acquiescence bias was verified, a new simulation study was conducted to determine the extent to which this absorption can affect the estimation of factor loadings. In this study, four different procedures were compared: a control group without any correction, acquiescence bias correction, residuals correction, and combined or mixed correction. Additionally, the study examined: (2) the magnitude of the factor loading estimates for the acquiescence bias factor, (3) the homogeneity of the estimated loadings, (4) the location of the residual doublet, and (5) the sign of items affected by the doublet.

To verify the potential effects due to the different procedures, the differences between the simulated and estimated factor loadings (in absolute value) of the two items involved in the simulated doublet were used as the dependent variable. For the control procedure, an exploratory factor analysis was conducted using the "*fa*" function from the "*psych*" package (Revelle, 2023). The acquiescence and residuals control (as well as the complete or mixed control) were carried out using the same methodology as the previous simulation.

So as to appraise whether there were differences in the estimation accuracy between the first and the second value of the doublet, a Student's-t test was conducted. The results indicated significant differences between the first and second items involved in the doublet ($t_{(63999)}$=30.441; p < .05). In Figure 3, a scatter plot is shown where the first and second elements that form the residual doublet are compared in pairs. A point with coordinates [-.2, -.1] would represent that in that replica, the difference between the simulated and estimated loading in the first element of the doublet is -.2, and the difference between the simulated and estimated value of the second element of the doublet is -.1. Overall, the trend followed by the scatterplot is directly proportional, meaning that the greater the bias (or difference between the estimated and simulated loading) in the first element of the doublet (V1), the greater the bias in the second element of the doublet (V2). However, if we consider the type of procedure used to analyze the data, it becomes evident that the trend changes. The control group appears to cluster into two clusters: one around values close to 0 and the other near values of .6. The residual group is characterized by having the highest variability in the results. The ACQ and mixed groups show a similar concentration of values near 0; however, there is a slight shift in the case of the mixed group, indicating that V1 values may exhibit greater bias than V2.

Figure 3.

Scatter plot that compares the estimated bias of the loadings of the items involved in the residual doublet.



The main results obtained from the two analyses of variance are in Table 4. The results of both are consistent, and agree in that the number of factors is the variable that has the greatest impact on the accuracy with which the loadings are estimated. This fact is reflected in a moderate effect size (for the second element or item of the doublet) and a high effect size (for the first item of the doublet). The impact of the correction is significant but with a small effect size for the first element of the doublet, while it does not appear to be non-trivial for the second element of the doublet.

Table 4.

Summary of ANOVAs.

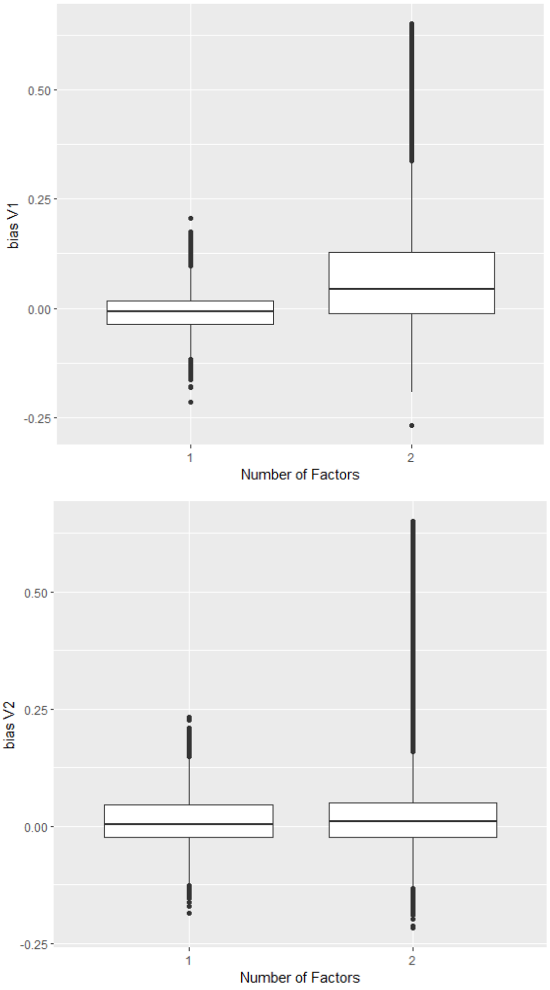
	V1 doublet				V2 doublet			
	F	Gl	p-value	Effect size	F	Gl	p-value	Effect size
Factor	11,296.325	1	< .05	.15*	2,746.91	1	< .05	.04
Sign	0	1	1	<.01	99.29	1	< .05	<.01
Correction	100,500.0	1	< .05	.02	2,129.82	3	< .05	<.01
Factor:Sign	0	1	1	<.01	28.16	1	< .05	<.01
Factor:Correction	895.0	1	< .05	<.01	954.46	3	< .05	.02
Sign:Correction	0	1	1	<.01	224.70	3	< .05	<.01
Factor:Sign:Correction	0	1	1	<.01	38.90	3	< .05	<.01

Note: V1 = first element of the doublet; V2 = second element of the doublet.

In Figure 4, displays two boxplots showing the accuracy in the estimation of loadings in V1 (left) and V2 (right), where the accuracy is compared when the residual is present between items within the same factor (Number of Factors = 1) and between items from different factors (Number of Factors = 2). In general, the medians and quartiles do not exhibit significant disparities among them; however, the greatest difference is observed in the extreme values. When the pair affects items within the same factor, the estimations are more accurate than those that involve the pair in two different factors. In other words, the difference between the simulated and estimated loading is more variable when the residuals are located in items belonging to different factors, both in V1 and V2.

Figure 4.

Box plot comparing the accuracy in item estimation when the pair is found between items of the same factor and between items from different factors.



Study 3: Empirical Example

So as to illustrate with real data the results obtaining via simulation, and assess the impact of controlling acquiescence in a dataset that presumably contains correlated residuals, we shall re-analyze an existing dataset used in the calibration of the Overall Personality Assessment Scales (OPERAS; Vigil-Colet et al., 2013). Specifically, we shall re-analyze the data corresponding to the Extraversion subscale (EX) of the Overall Personality Assessment Scales (OPERAS; Vigil-Colet et al., 2013). Respondents were 2,429 adults, with an age range between 18 and 60 years (M=29.15; SD = 14.65) and of which 38.37% were men. As for the EX subscale, it comprises 7 items that are almost fully balanced (4 measuring extraversion and 3 introversion), and all of them phrased positively. The scale scores are characterized by high reliability and very low levels of social desirability bias.

Four EFA's were conducted on this dataset: a content-only EFA without any correction, an EFA with acquiescence correction using the SIREN procedure (Navarro-Gonzalez et al., in press), an EFA with correlated-residuals correction using the MORGANA method (Ferrando et al., 2022), and a mixed exploratory analysis that corrected for acquiescence and removed correlated residuals. This last combined procedure was implemented in two steps: First, the SIREN procedure was used to partialize ACQ, and next, MORGANA factor analysis (Ferrando et al., in press) was fitted to the ACQ-partialized residual matrix.

Given the ordinal nature of the data (graded responses with less than 7 response options), the EFA's were based on polychoric item-item correlation matrices and fitted with the Unweighted Least Squares (ULS) criterion. All analyses were carried out using R, utilizing the same packages as in the previous simulation studies. The main results are in tables 5, 6, 7 and 8.

The most apparent result in table 5 is that the EREC index decreases dramatically when the data is pre-corrected for acquiescence. Without correction, residual correlation is detected between pairs 2 - 4 and 5 - 6 (see Table 5). The detected item pairs exhibit clear semantic redundancy. So, the detection results are submitted to be correct. However, once the variance due to acquiescence is partialled-out, no substantial correlated residuals are longer detected.

Table 5.

Detected doublets according to EREC index

Doublets	EREC Index	
	No correct	Correct
2 – 4	.578	
5 – 6	.470	.203
2 – 5		.120

Note. Values less than .2 will be considered trivial

Table 6.

Main detected doublets

Doublets	Items
2 – 4	2. Me desenvuelvo bien en situaciones sociales 4. Hago amigos con facilidad
5 – 6	5. Prefiero que otros sean el centro de atención 6. Permanezco en segundo plano

Note. 2. I handle social situations well; 4. I make friends easily; 5. I prefer others to be the center of attention; 6. I stay in the background

Table 7.

Estimated Loadings in Each of the Procedures.

Control	ACQ	Residuals	Combinated
660	.624	.685	.644
.766	.722	.696	.686
-.686	-.684	-.715	-.647
.746	.718	.674	.702
-.554	-.659	-.495	-.680
-.625	-.702	-.580	-.691
.664	.625	.688	.635

Table 7 compares the loading estimates when EFA's are performed (a) without correction, (b) correcting only acquiescence, (c) correcting only residuals, and (d) performing a complete correction. Factor loadings for items 5 and 6 exhibit the greatest variability across different correction types, the difference being maximal between the acquiescence correction option and the residuals correction option.

Table 8.

Goodness of Fit Indices

	GFI	TLI	RMSEA	RMSR
Control	---	.781	.17	.084
ACQ	.999	.998	.023	.021
Residuals	.987	.987	.047	---
Combined	.999	.999	.00	---

Table 8 displays the goodness of fit indices estimated in each of the procedures. Of the four, the control procedure is the only one that does not reach an acceptable fit. The ACQ correction and residual correction procedures exhibit good fit in GFI and TLI terms and moderate fit in RMSEA terms. As expected, the combined procedure yields the best results.

Discussion

The current research has attempted to explore the potential impact of two non-content sources of error or unwanted determinants of different origins: ACQ and correlated residuals. Previous studies in which both sources were considered separately found that both, ACQ and correlated residuals can distort structural estimation and goodness-of-fit assessment at the calibration step, and, individual score estimation at the scoring step. However, the combined effect of their joint occurrence does not appear to have been addressed until now.

Through three studies (two simulation studies and one empirical study), we have attempted to determine what occurs when we correct for ACQ in a dataset that includes more than one of the unwanted determinants. The predictions made at the beginning of this document provide analytical evidence that part of the correlated residual variance may be absorbed by the ACQ factor when ACQ corrections are applied, a prediction that has been supported by the results of the first simulation study. In this initial study, it was found that MORGANA, even when being a very sensitive procedure, was unable to detect almost any simulated residual doublet (true positive) when there was prior ACQ correction, regardless of the item's location (same factor or not) and the sign of the items (same sign, opposite sign). This result also holds in cases where two items correlate in the same factor but have different signs, as well as when the doublet is located in two different factors with no prior correction.

At the same time, however, the results suggest that, even though there is indeed a clear absorption effect by the ACQ factor, this effect does not seem to have a negative impact on the model fit results and the accuracy of the factor loading estimates corresponding to the content factors. In general, the trend when using the ACQ correction method and the combined method is that the accuracy in the estimation of content loadings is very good. Furthermore, the estimation of the second element of the pair is slightly more precise than that of the first element; however, the overall trend is that the greater the bias (difference between simulated and estimated loading) in the first element of the pair, the greater the bias will be in the second element of the pair.

It is important to stress that the present study only considered fully and essentially balanced item sets, and one-factor, and two-uncorrelated-factor models with high loadings on the content factors. A very simple an "ideal" set of conditions indeed. So, the results cannot be naively generalized to more complex models and further intensive research is needed on the present topics. However, even when acknowledging its preliminary nature, we believe that the results obtained here provide useful information that can be considered for practical applications.

Based on the results obtained here, it can be preliminarily concluded that, when correcting for acquiescence in a dataset that contains correlated residuals, we are absorbing not only the portion of variance attributed to this response bias but also part of the variance that is due to the existing correlated residuals. However, when both sources are jointly corrected, the absorption effect is expected to be much weaker, goodness of model-data fit is expected to slightly improve, and the structural estimates for the content factors are expected to be essentially unbiased. So, what we tentatively suggest is that, when fitting a balanced measure that already aims to correct for ACQ in a dataset in which correlated residuals are also suspected, the best approach is to perform a dual correction procedure.

References

Aichholzer, J. (2014). Random intercept EFA of personality scales. *Journal of Research in Personality, 53*, 1-4. https://doi.org/10.1016/j.jrp.2014.07.001

Baumgartner, H., & Steenkamp, J. B. E. (2001). Response styles in marketing research: A cross-national investigation. *Journal of marketing research, 38*(2), 143-156. https://doi.org/10.1509/jmkr.38.2.143.18840

Bentler, P. M., Jackson, D. N., & Messick, S. (1971). Identification of content and style: a two-dimensional interpretation of acquiescence. *Psychological Bulletin, 76*(3), 186-204. https://doi.org/10.1037/h0031474

Billiet, J. B & Davidov, E. (2008). Testing the stability of an acquiescence style factor behind two interrelated substantive variables in a panel design. *Sociological Methods Research, 36*(4), 542-562. https://doi.org/10.1177/0049124107313901

Billiet J. B. & McClendon M. K. (2000) Modeling Acquiescence in Measurement Models for Two Balanced Sets of Items. *Structural Equation Modeling: A Multidisciplinary Journal,* 7(4), 608-628. http://doi.org/1207/S15328007SEM0704_5

Chan, W., & Bentler, P. M. (1993). The covariance structure analysis of ipsative data. *Sociological Methods & Research, 22*(2), 214-247. https://doi.org/10.1177/0049124193022002003

Cheung, G. W., & Rensvold, R. B. (2000). Assessing extreme and acquiescence response sets in cross-cultural research using structural equations modeling. *Journal of cross-cultural psychology, 31*(2), 187-212. https://doi.org/10.1177/0022022100031002003

Cronbach, L. J. (1946). Response sets and test validity. *Educational and psychological measurement*, *6*(4), 475-494. https://doi.org/10.1177/001316444600600405

Danner, D., Aichholzer, J., & Rammstedt, B. (2015). Acquiescence in personality questionnaires: Relevance, domain specificity, and stability. *Journal of Research in Personality*, *57*, 119-130. https://doi.org/10.1016/j.jrp.2015.05.004

de la Fuente, J., & Abad, F. J. (2020). Comparing methods for modeling acquiescence in multidimensional partially balanced scales. *Psicothema 32*(4), 590-597. http://doi.org/10.7334/psicothema2020.96

Ferrando, P. J., & Anguiano-Carrasco, C. (2010). Acquiescence and social desirability as item response determinants: An IRT-based study with the Marlowe–Crowne and the EPQ Lie scales. *Personality and Individual Differences*, *48*(5), 596-600. https://doi.org/10.1016/j.paid.2009.12.013

Ferrando, P. J., Hernandez-Dorado, A., & Lorenzo-Seva, U. (2022). Detecting Correlated Residuals in Exploratory Factor Analysis: New Proposals and a Comparison of Procedures. *Structural Equation Modeling: A Multidisciplinary Journal, 29*(4), 630-638. https://doi.org/10.1080/10705511.2021.2004543

Ferrando, P. J., Hernandez-Dorado, A., & Lorenzo-Seva, U. (in press). A simple two-step procedure for fitting fully unrestricted exploratory factor analytic solutions with correlated residuals.

Ferrando, P. J., & Lorenzo-Seva, U. (2010). Acquiescence as a source of bias and model and person misfit: A theoretical and empirical analysis. *British Journal of Mathematical and Statistical Psychology, 63*(2), 427-448. https://doi.org/10.1348/000711009X470740

Ferrando, P. J., Lorenzo-Seva, U., & Chico, E. (2003). Unrestricted factor analytic procedures for assessing acquiescent responding in balanced, theoretically unidimensional personality scales. *Multivariate Behavioral Research, 38*(3), 353-374. https://doi.org/10.1207/S15327906MBR3803_04

Ferrando, P. J., Lorenzo-Seva, U., & Chico, E. (2009). A general factor-analytic procedure for assessing response bias in questionnaire measures. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(2), 364-381. https://doi.org/10.1080/10705510902751374

Hand, J., & Brazzell, C. O. (1965). Contamination in measures of acquiescence and social desirability. *Psychological Reports*, *16*(3), 759-760. https://doi.org/10.2466/pr0.1965.16.3.759

Hinz, A., Michalski, D., Schwarz, R., & Herzberg, P. Y. (2007). The acquiescence effect in responding to a questionnaire. *GMS Psycho-Social Medicine*, *4*, 1-9. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2736523/pdf/PSM-04-07.pdf

Lorenzo-Seva, U., & Ferrando, P.J. (2009). Acquiescent responding in partially balanced multidimensional scales. *British Journal of Mathematical and Statistical Psychology, 62*(2), 319-326. https://doi.org/10.1348/000711007X265164

Lorenzo-Seva, U., & Rodríguez-Fornells, A. (2006). Acquiescent responding in balanced multidimensional scales and exploratory factor analysis. *Psychometrika*, *71*(4), 769-777. https://doi.org/10.1007/s11336-004-1207-4

Maydeu-Olivares, A., & Coffman, D. L. (2006). Random intercept item factor analysis. *Psychological Methods, 11*(4), 344–362. https://doi.org/10.1037/1082-989X.11.4.344

Messick, S. (1995). Standards of validity and the validity of standards in performance asessment. *Educational measurement: Issues and practice*, *14*(4), 5-8. https://doi.org/10.1111/j.1745-3992.1995.tb00881.x

Montoya, A. K., & Edwards, M. C. (2021). The Poor Fit of Model Fit for Selecting Number of Factors in Exploratory Factor Analysis for Scale Evaluation. *Educational and Psychological Measurement, 81*(3), 413–440. https://doi.org/10.1177/0013164420942899

Morales-Vives, F., Cosi, S., Lorenzo-Seva, U., & Vigil-Colet, A. (2019). The inventory of callous-unemotional traits and antisocial behavior (INCA) for young people: Development and validation in a community sample. *Frontiers in Psychology, 10*:713. https://doi.org/10.3389/fpsyg.2019.00713

Navarro-González, D., Lorenzo-Seva, U., & Vigil-Colet, A. (2016). How response bias affects the factorial structure of personality self-reports. *Psicothema, 28*(4), 465-470. http://www.doi.org/10.7334/psicothema2016.113

Navarro-Gonzalez, D., Ferrando, P.J., Morales-Vives, F. & Hernandez-Dorado, A. (in press) SIREN: An Hybrid CFA-EFA R Package for Controlling Acquiescence in Restricted Factorial.

Park, M., & Wu, A. D. (2019). Item Response Tree Models to Investigate Acquiescence and Extreme Response Styles in Likert-Type Rating Scales. *Educational and Psychological Measurement, 79*(5), 911–930. https://doi.org/10.1177/0013164419829855

Podsakoff, P.M, MacKenzie, S.B., Lee, J., & Podsakoff, N.P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *885*(879), 1010-1037. https://doi.org/10.1037/0021-9010.88.5.879

Primi, R., Santos, D., De Fruyt, F., & John, O. P. (2019). Comparison of classical and modern methods for measuring and correcting for acquiescence. *British Journal of Mathematical and Statistical Psychology*, *72*(3), 447-465. https://doi.org/10.1111/bmsp.12168

Ray, J. J. (1979). Is the acquiescent response style not so mythical after all? Some results from a successful balanced F scale. *Journal of Personality Assessment, 43*(6), 638–643. https://doi.org/10.1207/s15327752jpa4306_14

Ray, J. J. (1983). Reviving the problem of acquiescent response bias. *Journal of Social Psychology, 121*(1), 81–96. https://doi.org/10.1080/00224545.1983.9924470

Revelle, M. W. (2015). Package 'psych'. *The comprehensive R archive network*, *337*(338). https://mirror.ibcp.fr/pub/CRAN/web/packages/psych/psych.pdf

Ross, C. E., & Mirowsky, J. (1984). Socially-desirable response and acquiescence in a cross-cultural survey of mental health. *Journal of Health and Social Behavior*,*25*(2), 189-197. https://doi.org/10.2307/2136668

Saris, W. E., Satorra, A., & Sörbom, D. (1987). The detection and correction of specification errors in structural equation models. *Sociological methodology*, 105-129. https://doi.org/10.2307/271030

Savalei, V., & Falk, C. F. (2014). Recovering substantive factor loadings in the presence of acquiescence bias: A comparison of three approaches. *Multivariate behavioral research*, *49*(5), 407-424. https://doi.org/10.1080/00273171.2014.931800

Vigil-Colet, A., Morales-Vives, F., Camps, E., Tous, J., & Lorenzo-Seva, U. (2013). Development and validation of the overall personality assessment scale (OPERAS). *Psicothema*, *25*(1), 100-106. https://www.redalyc.org/pdf/727/72725690017.pdf

Weijters, B.; Geuens, M.; Schillewaert, N. (2010). The Individual Consistency of Acquiescence and Extreme Response Style in Self-Report Questionnaires. *Applied Psychological Measurement*, *34*(2), 105 -121. https://doi.org/10.1177/0146621609338593

Capítulo 5. Conclusiones

So far, an extensive amount of literature associated with Acquiescence (ACQ) research is available. Part of this research has focused on identifying which variables – personality, situational, or item-related – are the main determinants of acquiescence (e.g., Wetzel et al., 2015), while others have been more focused on the impact of acquiescent responding on the factorial structures of personality measures and goodness of model-data fit (Navarro-Gonzalez et al., 2016). There is also research about methods for detecting and correcting bias, with perhaps the most widely used and effective method being item balancing (Baumgartner & Steenkamp, 2001; Nunnally, 1978). However, this method requires specific characteristics (using antonyms instead of negatives, ensuring that antonyms are not more abstract or difficult to understand than positives) to be met so as to be truly effective. Achieving these characteristics can be challenging when designing a test, and in many cases, post-hoc procedures like RIFA (Maydeu-Olivares & Coffman, 2006) or RI-EFA (Aichholzer, 2014), among others, are necessary for correcting the impact of acquiescence. The effectiveness of these commonly used methods has been demonstrated by authors such as Savalei & Falk (2014) and de la Fuente & Abad (2020).

In contrast to the evidence summarized above, there does not seem to be much interest in assessing the impact that the acquiescence control methods have on external validity relations, despite external validity being considered an indicator of measurement quality. (Jenkinson et al., 1994). Furthermore, researchers seem to prefer to focus on studying the impact of unwanted determinants on a one-by-one basis, and the study of the combined impact of various sources of error seems to have taken a back seat.

Acknowledging this gap in the literature, the present research has aimed to study the impact of ACQ on external validity relations and assess the effect of controlling this bias when at least two determinants unrelated to content are presented together. To achieve these two general aims in this thesis, five articles have been completed, and two procedures have been developed (a hybrid method for detecting and correcting ACQ and another for detecting correlated residuals).

To assess the effect of post hoc acquiescence correction on external validity, we used the R implementation of the Partially Balanced EFA procedure (Lorenzo-Seva & Ferrando, 2009), which is available in the vampyr package (Navarro-Gonzalez, 2020). The results revealed that, as expected, the empirical validity estimate is an attenuated (downwardly biased) estimate of theoretical validity, and that the amount of discrepancy (i.e. the amount of bias) mostly depends on the internal characteristics of the test. When an ACQ correction based on balanced scales is used, the discrepancy between estimated and theoretical validity decreases slightly but significantly. However, in scenarios in which the extraction of variance due to ACQ is more challenging (low loadings on the ACQ factor with a highly heterogeneous loading pattern) and/or the test has clearly unfavorable characteristics (longer, more abstract, or harder-to-understand antonymous items), the difference between theoretical and empirical validity will be greater. It can be concluded that, although post hoc control is appropriate and leads to improvements in the behavior of the empirical validity as an estimate, the most critical determinant in the accuracy of external validity is the appropriate selection of balanced items.

As for the second main purpose, the development of the SIREN and MORGANA methods and their subsequent implementation in R allowed us to assess the effect of ACQ correction when more than one determinant unrelated to the construct is operating. The SIREN method is a hybrid ACQ control procedure (combining EFA and CFA) that operates sequentially in three steps: In the first stage, the acquiescence factor is identified, and its effects are partially removed from the item correlations. In the second stage, a specified confirmatory solution is either fitted to the "cleaned" reduced correlation matrix to the initial matrix but specifying a complete solution with fixed loadings on the ACQ factor. In the third stage, finally factor scores are estimated for each individual. According to the simulation results, the functioning of the procedure is quite stable. Furthermore, it can be considered to be more flexible and applicable to a wider range of situations than existing related procedures, as this method allows for obtaining clean individual score estimates, and its hybrid nature provides flexibility in the estimation and fitting procedures.

With regards to MORGANA, it was developed in response to the need for efficient detection of correlated doublets, as existing approaches at the time (the standard residual covariance matrix method and the anti-image partial correlation matrix method) overlooked the potential propagation effect of doublets when the residual covariance matrix is constrained to be identity. In effect, in a simulation study, it was observed that a significant portion of doublets propagated, leading the unmodeled covariance to be reassigned in the factor loadings or spread across the entire residual covariance matrix, thus making their detection with available methods quite challenging. The two indices implemented in the MORGANA procedure are based on previous contributions. In simulation studies, EREC proved to be the most effective and accurate diagnostic index, especially under conditions of large samples, many items per factor, low communality, and a high value of doublets. Based on these indices, an extended EFA solution was next proposed in which a residual covariance structure was incorporated and there was no need to specify a priori which were the residual elements that have to be estimated.

Used separately, both SIREN and MORGANA have proven to useful tools, as they are effective, accurate, and provide a more robust alternative to the limitations of other competing approaches. In the complex and noisy datasets that are found in real-life applications, the combined use of these two procedures significantly improves the accuracy of the structural estimates, model-data fit assessment and estimated content scores.

Continuing with the development of these procedures and their implementation in R, research was undertaken to achieve the second primary objective of the present work: to explore the combined impact of two sources of error unrelated to content that, individually, are potentially capable of distorting parameter estimation and fit assessment. In spite of its relevance and high probability of occurrence, this topic has received very little interest in the literature.

In the present study, the combined impact of ACQ and doublets was evaluated through two simulation studies and an empirical study, and the most remarkable observed results were: (a) SIREN tends to correct both the variance attributed to response bias and a significant portion of the variance due to residual correlations. (b) a tendency for MORGANA to over-detect doublets, which could be due to the dispersion of the unmodeled residual variance throughout the entire residual covariance matrix, and (c) good model-data fit results when the two correction methods were used in combined form.

Given that, available instruments with "ideal" properties are scarce and that it is not always possible to design them, the results obtained here provide reasonably clear information regarding the need for post-hoc response bias control and/or residual detection techniques, as, ultimately, an improvement in validity is expected. In addition, even when he bias correction procedure might "absorb" the unwanted variance without establishing a clear delimitation (between ACQ and correlated specificities), this lack of separation does not harm both the accuracy and estimation of the content parameters or the model-data fit assessment results. Now, it would be interesting to continue researching in this line and assess the extent to which this absorption effect would generalize to other response styles or response bias response bias that tend to be modeled in a similar manner.

At present, and for large and complex models, it is virtually impossible to predict what kind of propagation (if any) would have a residual, which may go to: (a) the factor loadings, (b) the inter-factor correlations or (c) be dispersed throughout the residual correlation matrix. However, the combined use of the two techniques considered here may be the most appropriate alternative to avoid the propagation effect to occur or to control them. On one hand, the EREC index is very sensitive, so those doublets that could be considered false positives (detected doublets that don't actually exist) could in fact be part of the non modeled residual covariance that has propagated throughout the entire residual matrix. On the other hand, the ENIDE index could detect doublets that have been reassigned to the factor loadings; however, if this residual variance were absorbed by the acquiescence factor, the bias correction method would have then corrected not only ACQ but also part of the variance associated with the non modeled residual.

In conclusion, at the substantive level, evidence has been provided for an improvement in external validity when ACQ bias is corrected, and, at the instrumental level, two techniques for detecting and correcting unwanted determinants have been further developed and implemented in R. Finally, a combined procedure in which both methods are used has been proposed for scenarios in which the simultaneous presence of residuals and ACQ is suspected.

Bibliografía

Aichholzer, J. (2014). Random intercept EFA of personality scales. *Journal of research in personality*, *53*, 1-4. https://doi.org/10.1016/j.jrp.2014.07.001

Angleitner, A., John, O. P., & Löhr, F. J. (1986). It's what you ask and how you ask it: An itemmetric analysis of personality questionnaires. In *Personality assessment via questionnaires: Current issues in theory and measurement* (pp. 61-108).Springer Berlin Heidelberg.

Bachman, J. G., & O'Malley, P. M. (1984). Yea-saying, nay-saying, and going to extremes: Black-white differences in response styles. *Public Opinion Quarterly*, *48*(2), 491-509. https://doi.org/10.1086/268845

Bandalos, D. L. (2021). Item meaning and order as causes of correlated residuals in confirmatory factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(6), 903-913. https://doi.org/10.1080/10705511.2021.1916395

Baumgartner, H., & Steenkamp, J. E. (2001). Response Styles in Marketing Research: A Cross-National Investigation. *Journal of Marketing Research, 38*(2), 143-156, https://doi.org/10.1509/jmkr.38.2.143.18840

Bentler, P. M., Jackson, D. N., & Messick, S. (1971). Identification of content and style: A two-dimensional interpretation of acquiescence. *Psychological Bulletin*, *76*(3), 186–204. https://doi.org/10.1037/h0031474

Billiet, J. B & Davidov, E. (2008). Testing the stability of an acquiescence style factor behind two interrelated substantive variables in a panel design. *Sociological Methods Research, 36*(4), 542-562. https://doi.org/10.1177/0049124107313901

Billiet, J. B., & McClendon, M. J. (2000). Modeling acquiescence in measurement models for two balanced sets of items. *Structural equation modeling*, *7*(4), 608-628, https://doi.org/10.1207/S15328007SEM0704_5

Cambré, B., Welkenhuysen-Gybels, J., & Billiet, J. (2002). Is it content or style? An evaluation of two competitive measurement models applied to a balanced set of ethnocentrism items. *International*

Journal of Comparative Sociology, *43*(1), 1-20. https://doi.org/10.1177/002071520204300101

Cattell, R. B. (1944). Psychological measurement: normative, ipsative, interactive. *Psychological Review, 51*(5), 292–303. https://doi.org/10.1037/h0057299

Chan, W., & Bentler, P. M. (1993). The covariance structure analysis of ipsative data. *Sociological Methods & Research*, *22*(2), 214-247. https://doi.org/10.1177/0049124193022002003

Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*(3), 265-289. https://doi.org/10.3102/10769986022003265

Chico, E., & Lorenzo-Seva, U. (2006). Belief in astrology inventory: Development and validation. *Psychological Reports*, *99*(3), 851-863. https://doi.org/10.2466/PR0.99.3.851-863

Clemans, W. V. (1968). Interest measurement and the concept of ipsativity. *Measurement and Evaluation in Guidance*, *1*(1), 50-55.

Cochran, W. G. (1954). Some methods for strengthening the common χ 2 tests. *Biometrics*, *10*(4), 417-451. https://doi.org/10.2307/3001616

Cole, D. A., Ciesla, J. A., & Steiger, J. H. (2007). The insidious effects of failing to include design-driven correlated residuals in latent-variable covariance structure analysis. *Psychological Methods, 12*(4), 381–398. https://doi.org/10.1037/1082-989X.12.4.381

Condon, L., Ferrando, P. J., & Demestre, J. (2006). A note on some item characteristics related to acquiescent responding. *Personality and individual differences*, *40*(3), 403-407. https://doi.org/10.1016/j.paid.2005.07.019

Couch, A., & Keniston, K. (1960). Yeasayers and naysayers: Agreeing response set as a personality variable. *The Journal of Abnormal and Social Psychology, 60*(2), 151–174. https://doi.org/10.1037/h0040372

Couch, A., & Keniston, K. (1961). Agreeing response set and social desirability. *The Journal of Abnormal and Social Psychology, 62*(1), 175–179. https://doi.org/10.1037/h0047429

Cordero, A., Seisdedos, N., González, M., & de la Cruz, V. (1989). *PMA. Aptitudes Primarias Mentales* [Primary Mental Abilities]. TEA Ediciones.

Cronbach, L. J. (1946). Response sets and test validity. *Educational and psychological measurement*, *6*(4), 475-494. https://doi.org/10.1177/001316444600600405

Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and psychological measurement*, *64*(3), 391-418. https://doi.org/10.1177/0013164404266386

Cruse, D. B. (1966). Some relations between minimal content, acquiescent-dissentient, and social desirability scales. *Journal of Personality and Social Psychology*, *3*(1), 112–119. https://doi.org/10.1037/h0022745

de la Fuente, J., & Abad, F. J. (2020). Comparing Methods for Modeling Acquiescence in Multidimensional Partially Balanced Scales. *Psicothema*, *32*(4), 590-597. http://10.7334/psicothema2020.96

DeMars, C. E. (2021). Violation of Conditional Independence in the Many-Facets Rasch Model. *Applied Measurement in Education*, *34*(2), 122-138. https://doi.org/10.1080/08957347.2021.1890743

DiStefano, C., & Motl, R. W. (2009). Self-esteem and method effects associated with negatively worded items: Investigating factorial invariance by sex. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(1), 134-146. https://doi.org/10.1080/10705510802565403

Dolnicar, S., & Grün, B. (2007). Cross-cultural differences in survey response patterns. *International Marketing Review*, *24*(2), 127-143. https://doi.org/10.1108/02651330710741785

Ebesutani, C., Drescher, C. F., Reise, S. P., Heiden, L., Hight, T. L., Damon, J. D., & Young, J. (2012). The importance of modeling method effects: Resolving the (uni) dimensionality of the loneliness questionnaire. *Journal of Personality Assessment*, *94*(2), 186-195. https://doi.org/10.1080/00223891.2011.627967

Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Psychology Press.

Ferrando, P. J., & Lorenzo ‐ Seva, U. (2010). Acquiescence as a source of bias and model and person misfit: A theoretical and empirical analysis. *British Journal of Mathematical and Statistical Psychology*, *63*(2), 427-448. https://doi.org/10.1348/000711009X470740

Ferrando, P. J., Lorenzo-Seva, U., & Chico, E. (2003). Unrestricted factor analytic procedures for assessing acquiescent responding in balanced, theoretically unidimensional personality scales. *Multivariate Behavioral Research, 38*(3), 353-374, https://doi.org/10.1207/S15327906MBR3803_04

Ferrando, P. J., Lorenzo-Seva, U., & Chico, E. (2009). A general factor-analytic procedure for assessing response bias in questionnaire measures. Structural Equation Modeling: *A Multidisciplinary Journal, 16*(2), 364-381. https://doi.org/10.1080/10705510902751374

Fischer, R. (2004). Standardization to account for cross-cultural response bias: A classification of score adjustment procedures and review of research in JCCP. *Journal of Cross-Cultural Psychology*, *35*(3), 263-282. https://doi.org/10.1177/0022022104264122

Forehand, G. A. (1962). Relationships among response sets and cognitive behaviors. *Educational and Psychological Measurement*, *22*(2), 287-302. https://doi.org/10.1177/001316446202200204

Fornell, C. (1983). Issues in the application of covariance structure analysis: A comment. *Journal of consumer research*, *9*(4), 443-448. https://www.jstor.org/stable/2488794

Gilbert, D.T. (1991). How mental systems believe. *American Psychologist, 46*, 107-119. https://doi.org/10.1037/0003-066X.46.2.107

Green, S. B., & Hershberger, S. L. (2000). Correlated errors in true score models and their effect on coefficient alpha. *Structural equation modeling*, *7*(2), 251-270. https://doi.org/10.1207/S15328007SEM0702_6

Gudjonsson, G. H. (1990). The relationship of intellectual skills to suggestibility, compliance and acquiescence. *Personality and individual differences*, *11*(3), 227-231. https://doi.org/10.1016/0191-8869(90)90236-K

Hanley, C. (1961). Social desirability and response bias in the MMPI. *Journal of Consulting Psychology*, *25*(1), 13 –20. https://doi.org/10.1037/h0043696

Horan, P. M., DiStefano, C., & Motl, R. W. (2003). Wording effects in self-esteem scales: Methodological artifact or response style?. *Structural Equation Modeling*, *10*(3), 435-455. https://doi.org/10.1207/S15328007SEM1003_6

Ip, E. H. S. (2001). Testing for local dependency in dichotomous and polytomous item response models. *Psychometrika*, *66*(1), 109-132. https://doi.org/10.1007/BF02295736

Jenkinson, C., Wright, L., & Coulter, A. (1994). Criterion validity and reliability of the SF-36 in a population sample. *Quality of Life Research*, *3*, 7-12. https://doi.org/10.1007/BF00647843

Johanson, G. A., & Osborn, C. J. (2004). Acquiescence as differential person functioning. *Assessment & Evaluation in Higher Education*, *29*(5), 535-548. https://doi.org/10.1080/02602930410001689126

Knowles, E. S., & Condon, C. A. (1999). Why people say" yes": A dual-process theory of acquiescence. *Journal of Personality and Social Psychology*, *77*(2), 379 –386. https://doi.org/10.1037/0022-3514.77.2.379

Kolenikov, S. (2011). Biases of parameter estimates in misspecified structural equation models. *Sociological methodology*, *41*(1), 119-157. https://doi.org/10.1111/j.1467-9531.2011.01236.x

Krosnick, J. A. (1999) Maximizing questionnaire quality, in: J. P. Robinson, P. R. Shaver & L.S. Wrightsman (Eds) *Measures of political attitudes* (San Diego, CA, Academic Press), 37–57.

Krosnick, J. A., Judd, C. M., & Wittenbrink, B. (2005). Attitude measurement. *Handbook of attitudes and attitude change. Mahwah, NJ: Erlbaum*, 21-76. https://web.stanford.edu/dept/communication/faculty/krosnick/docs/2006/2006%20Attitude%20Measurement%20Techniques%20for%20Measuring%20the%20Unobservab.pdf

Krosnick, J. A., & Presser, S. (2010). Handbook of survey research: Question and questionnaire design. *Handbook of survey research: Question and Questionnaire Design*, *2*, 264-313.

Leary, M. R., & Kowalski, R. M. (1990). Impression management: A literature review and two-component model. *Psychological bulletin*, *107*(1), 34-47. https://doi.org/10.1037/0033-2909.107.1.34

Lord, F. M., & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores, Reading*. MA:Addison-Wesley.

Lorenzo-Seva, U., & Ferrando, P. J. (2009). Acquiescent responding in partially balanced multidimensional scales. *British Journal of*

Mathematical and Statistical Psychology, *62*(2), 319-326. https://doi.org/10.1348/000711007X265164

Lucke, J. F. (2005). "Rassling the hog": The influence of correlated item error on internal consistency, classical reliability, and congeneric reliability. *Applied psychological measurement*, *29*(2), 106-125. https://doi.org/10.1177/0146621604272739

Malhotra, N., & Krosnick, J. A. (2007). The effect of survey mode and sampling on inferences about political attitudes and behavior: Comparing the 2000 and 2004 ANES to Internet surveys with nonprobability samples. *Political Analysis*, *15*(3), 286-323. https://doi.org/10.1093/pan/mpm003

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the national cancer institute*, *22*(4), 719-748. https://doi.org/10.1093/jnci/22.4.719

Marais, I., & Andrich, D. (2008). Effects of Varying Magnitude and Patterns of Response Dependence. *Journal of applied measurement*, *9*(2), 105-124. http://publicifsv.sund.ku.dk/~kach/PsyLab2018/Marais,%20Andrich,%202008.pdf

Maydeu-Olivares, A., & Coffman, D. L. (2006). Random intercept item factor analysis. Psychological Methods, 11(4), 344–362. https://doi.org/10.1037/1082-989X.11.4.344

Meisenberg, G., & Williams, A. (2008). Are acquiescent and extreme response styles related to low intelligence and education?. *Personality and individual differences*, *44*(7), 1539-1550. https://doi.org/10.1016/j.paid.2008.01.010" \t "_blank" \o "Persistent link using digital object identifier

Messick, S. (1990). Validity of test interpretation and use.

Mirowsky, J., & Ross, C. E. (1991). Eliminating defense and agreement bias from measures of the sense of control: A 2 x 2 index. Social Psychology Quarterly, 127-145. https://doi.org/10.2307/2786931

Morales-Vives, F., Camps, E., & Lorenzo-Seva, U. (2013). Development and validation of the psychological maturity assessment scale (PSYMAS). *European Journal of Psychological Assessment*. https://doi.org/10.1027/1015-5759/a000115

Morales-Vives, F., Gómez-Herrera, M., & Vigil-Colet, A. (2020). INCA-M: Mexican Adaptation of the Inventory of Callous-Unemotional Traits and Antisocial Behavior. *Frontiers in psychology*, *11*, 753. https://doi.org/10.3389/fpsyg.2020.00753

Morales-Vives, F., Lorenzo-Seva, U., & Vigil-Colet, A. (2017). Cómo afectan los sesgos de respuesta a la estructura factorial de los tests basados en el modelo de los Cinco Grandes factores de personalidad. [How response biases affect the factor structure of Big Five personality questionnaires]. *Anales De Psicología/Annals of Psychology*, *33*(3), 589-596. https://doi.org/10.6018/analesps.33.3.254841

Navarro-González, D., Ferrando, P. J., & Vigil-Colet, A. (2018). Is general intelligence responsible for differences in individual reliability in personality measures?. *Personality and Individual Differences*, *130,* 1-5. https://doi.org/10.1016/j.paid.2018.03.034

Navarro-Gonzalez, D., Lorenzo-Seva, U., & Vigil-Colet, A. (2016). How response bias affects the factorial structure of personality self-reports. *Psicothema, 28*(4), 465-470. https://doi.org/10.7334/psicothema2016.113

Nunnally, J.C. (1978) Psychometric theory. 2nd Edition, McGraw-Hill.

Pasek, J., & Krosnick, J. A. (2010). Measuring intent to participate and participation in the 2010 census and their correlates and trends: Comparisons of RDD telephone and non-probability sample Internet survey data. *Survey Methodology*, *2010,* 15.

https://www.census.gov/content/dam/Census/library/working-papers/2010/adrm/ssm2010-15.pdf

Paulhus, D. L., & Vazire, S. (2007). The self-report method. *Handbook of research methods in personality psychology, 1*(2007), 224-239.

Primi, R., De Fruyt, F., Santos, D., Antonoplis, S., & John O. P. (2019). True or False? Keying Direction and Acquiescence Influence the Validity of Socio-Emotional Skills Items in Predicting High School Achievement. *International Journal of Testing 20*(2), 97-121. https://doi.org/10.1080/15305058.2019.1673398

Primi, R., Santos, D., De Fruyt, F., & John, O. P. (2019). Comparison of classical and modern methods for measuring and correcting for acquiescence. *British Journal of Mathematical and Statistical Psychology, 72*(3), 447-465. https://doi.org/10.1111/bmsp.12168

Raven, J. C. (1996). *Matrices progresivas. Escalas CPM Color y SPM General*. [Raven Progressive Matrices]. TEA Ediciones.

Ray, J. J. (1983). Reviving the problem of acquiescent response bias. *The Journal of Social Psychology, 121*(1), 81-96, http://doi.org/10.1080/00224545.1983.9924470

Revelle, W. (2021). *Psych: Procedures for psychological, psychometric, and personality research* (version 2.1.6) [R package]. https://cran.rstudio.org/web/packages/psych/psych.pdf

Ross, C. E., & Mirowsky, J. (1984). Socially-desirable response and acquiescence in a cross-cultural survey of mental health. *Journal of Health and Social Behavior*, 189-197. https://doi.org/10.2307/2136668

Ruiz-Pamies, M., Lorenzo-Seva, U., Morales-Vives, F., Cosi, S., & Vigil-Colet, A. (2014). I-DAQ: a new test to assess direct and indirect aggression free of response bias. *The Spanish Journal of Psychology, 17*, E41. https://doi.org/10.1017/sjp.2014.43

Rundquist, E. A. (1966). Item and response characteristics in attitude and personality measurement: A reaction to LG Rorer's" The great response-style myth.". *Psychological Bulletin*, *66*(3), 166 –177. https://doi.org/10.1037/h0023709

Saris, W. E., & Aalberts, C. (2003). Different explanations for correlated disturbance terms in MTMM studies. *Structural Equation Modeling*, *10*(2), 193-213. https://doi.org/10.1207/S15328007SEM1002_2

Savalei, V., & Falk, C. F. (2014). Recovering substantive factor loadings in the presence of acquiescence bias: A comparison of three approaches. *Multivariate behavioral research*, *49*(5), 407-424. https://doi.org/10.1080/00273171.2014.931800

Scharl, A., & Gnambs, T. (2022). The impact of different methods to correct for response styles on the external validity of self-reports. *European Journal of Psychological Assessment*. https://doi.org/10.1027/1015-5759/a000731

Schlenker, B. R. (1980). *Impression management* (Vol. 526). Monterey, CA: Brooks/Cole.

Schneider, D. J. (1981). Toward a Broader Conception. En J. T. Tedeschi (Ed.), *Impression management theory and social psychological research* (pp. 23-40). Academic Press.

Schriesheim, C. A., & Hill, K. D. (1981). Controlling acquiescence response bias by item reversals: The effect on questionnaire validity. *Educational and psychological measurement*, *41*(4), 1101-1114. https://doi.org/10.1177/001316448104100420

Shi, G., Huang, T., Dong, W., Wu, J., & Xie, X. (2018). Robust foreground estimation via structured Gaussian scale mixture modeling. *IEEE Transactions on Image Processing*, *27*(10), 4810-4824. https://doi.org/10.1109/TIP.2018.2845123

Sörbom, D. (1989). Model modification. *Psychometrika*, *54*(3), 371-384. https://doi.org/10.1007/BF02294623

Soto, C. J., & John, O. P. (2019). Optimizing the length, width, and balance of a personality scale: How do internal characteristics affect external validity?. *Psychological assessment*, *31*(4), 444-459. https://doi.org/10.1037/pas0000586

Soto, C. J., John, O. P., Gosling, S. D., & Potter, J. (2008). The developmental psychometrics of Big Five self-reports: Acquiescence, factor structure, coherence, and differentiation from ages 10 to 20. *Journal of Personality and Social Psychology, 94,* 718-737. doi: 10.1037/0022-3514.94.4.718

Soto, C. J., John, O. P., Gosling, S. D., & Potter, J. (2011). Age differences in personality traits from 10 to 65: Big five domains and facets in a large cross-sectional sample. *Journal of Personality and Social Psychology,100,* 330-348. doi: 10.1037/a0021717

Taylor, J. C., & Bowers, D. G. (1972). Survey of organizations: A machine-scored standardized questionnaire instrument. U. Michigan

ten Berge, J. M. F. (1999). A legitimate case of component analysis of ipsative measures, and partialling the mean as an alternative to ipsatization. *Multivariate Behavioral Research*, *34*(1), 89–

102. https://doi.org/10.1207/s15327906mbr3401_4

Thissen, D., Bender, R., Chen, W., Hayashi, K., & Wiesen, C. A. (1992). Item response theory and local dependence: A preliminary report (Research Memorandum 92-2). *Chapel Hill: LL Thurstone Laboratory, University of North Carolina at Chapel Hill*.

Tourangeau, R., Rasinski, K. A., Bradburn, N., & D'Andrade, R. (1989). Belief accessibility and context effects in attitude measurement. *Journal of Experimental Social Psychology*, *25*(5), 401-421. https://doi.org/10.1016/0022-1031(89)90030-9

Van Herk, H., Poortinga, Y. H., & Verhallen, T. M. (2004). Response styles in rating scales: Evidence of method bias in data from six EU countries. *Journal of Cross-Cultural Psychology*, *35*(3), 346-360. https://doi.org/10.1177/0022022104264126

Vigil-Colet, A., Lorenzo-Seva, U., & Condon, L. (2008). Development and validation of the statistical anxiety scale. *Psicothema, 20*(1), 174-180. http://www.psicothema.com/pdf/3444.pdf

Vigil-Colet, A., Lorenzo-Seva, U., & Morales-Vives, F. (2015). The effects of ageing on self-reported aggression measures are partly explained by response bias. *Psicothema. 27*(3), 209-215. doi: 10.7334/psicothema2015.32

Vigil-Colet, A., Morales-Vives, F., & Lorenzo-Seva, U. (2013). How social desirability and acquiescence affect the age-personality relationship. *Psicothema*, *25*(3), 342-348. https://www.redalyc.org/pdf/727/72728043010.pdf

Vigil-Colet, A., Navarro-González, D., & Morales-Vives, F. (2020). To reverse or to not reverse Likert-type items: That is the question. *Psicothema*, *32*(1), 108–114. https://doi.org/10.7334/psicothema2019.286

Watson, D. (1992). Correcting for acquiescent response bias in the absence of a balanced scale: An application to class consciousness. *Sociological Methods & Research*, *21*(1), 52-88. https://doi.org/10.1177/0049124192021001003

Webster, H. (1958). Correcting personality scales for response sets or suppression effects. *Psychological Bulletin*, *55*(1), 62–64. https://doi.org/10.1037/h0048031

Wechsler, D. (2003). *Escala de inteligencia de Wechsler para niños-IV (WISC-IV)* [Wechsler Intelligence Scale for Children-WISC-IV]. Psychological Corporation.

Weijters, B., & Baumgartner, H. (2022). On the use of balanced item parceling to counter acquiescence bias in structural equation models. *Organizational Research Methods*, *25*(1), 170-180. https://doi.org/10.1177/1094428121991909

Weijters, B., Geuens, M., & Schillewaert, N. (2010a). The individual consistency of acquiescence and extreme response style in self-report questionnaires. *Applied Psychological Measurement, 34*(2), 105-121. doi: 10.1177/0146621609338593

Weijters, B., Geuens, M., & Schillewaert, N. (2010b). The stability of individual response styles. *Psychol Methods, 15*(1), 96-110. doi: 10.1037/a0018721

Welkenhuysen-Gybels, J., Billiet, J., & Cambré, B. (2003). Adjustment for Acquiescence in the Assessment of the Construct Equivalence of Likert-Type Score Items. *Journal of Cross-Cultural Psychology, 34*(6), 702–722. https://doi.org/10.1177/0022022103257070

Wetzel, E., Böhnke, J. R. & Brown, A., (2016) Response biases. In F. T. L. Leong, D. Bartram, F. Cheung, K. F. Geisinger, & D. Iliescu (Eds.), *The ITC International Handbook of Testing and Assessment* (pp. 349-363). Oxford University Press.

Winkler, J. D., Kanouse, D. E., & Ware, J. E. (1982). Controlling for Acquiescence Response Set in scale development. *Journal of Applied Psychology*, *67*(5), 555 –561. https://doi.org/10.1037/0021-

9010.67.5.555

Wong, N., Rindfleisch, A., & Burroughs, J. E. (2003). Do reverse-worded items confound measures in cross-cultural consumer research? The case of the material values scale. *Journal of consumer research, 30*(1), 72-91. https://doi.org/10.1086/374697

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8*(2), 125-145. https://doi.org/10.1177/014662168400800201

Zhang, L., Pan, J., Dubé, L., & Ip, E. H. (2021). blcfa: An R package for Bayesian model modification in confirmatory factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(4), 649-658. https://doi.org/10.1080/10705511.2020.1867862

Zuckerman, M., Knee, C.R., Miyake, K. & Hodgingd, H.S. (1995). Hypothesis Confirmation: The Joint Effect of

Positive Test Strategy and Acquiescence Response Set.

Journal of Personality and Social Psychology, *68*, 52-60.