Psychological Medicine

www.cambridge.org/psm

Do complex psychometric analyses really matter? Comparing multiple approaches using individual participant data from antidepressant trials

Original Article

Cite this article: Byrne, D., Doyle, F., Brannick, S., Carney, R. M., Cuijpers, P., Dima, A. L., Freedland, K. E., Guerin, S., Hevey, D., Kathuria, B., Wallace, E., & Boland, F. (2025). Do complex psychometric analyses really matter? Comparing multiple approaches using individual participant data from antidepressant trials. *Psychological Medicine*, **55**, e289, 1–9
https://doi.org/10.1017/S0033291725101785

Received: 04 February 2025
Revised: 21 August 2025
Accepted: 26 August 2025

Keywords:
antidepressant agent; depressive disorder; psychiatric status rating scales; randomized controlled trails; psychometrics

Corresponding author:
David Byrne;
Email: david.byrne@rcsi.ie

David Byrne[1], Frank Doyle[2], Susan Brannick[3], Robert M Carney[4], Pim Cuijpers[5], Alexandra L Dima[6,7,8], Kenneth E Freedland[4], Suzanne Guerin[9], David Hevey[10], Bishember Kathuria[11], Emma Wallace[12,13] and Fiona Boland[2]

[1]School of Population Health, Dublin, Ireland; [2]School of Population Health, RCSI University of Medicine and Health Sciences, Dublin, Ireland; [3]Aware, Dublin, Ireland; [4]Department of Psychiatry, Washington University School of Medicine, St Louis, USA; [5]Department of Clinical, Neuro and Developmental Psychology, Vrije Universiteit Amsterdam. Amsterdam, the Netherlands; [6]Avedis Donabedian Research Institute, Autonomous University of Barcelona, Barcelona, Spain; [7]Health Technology Assessment in Primary Care and Mental Health (PRISMA), Institut de Recerca Sant Joan de Déu, Esplugues de Llobregat, Spain; [8]Consortium "Centro de Investigación Biomédica en Red" Epidemiology and Public Health (CIBERESP), Madrid, Spain; [9]School of Psychology, University College Dublin. Dublin, Ireland; [10]School of Psychology, Trinity College Dublin. Dublin, Ireland; [11]Medical Affairs, Novartis Ireland Ltd., Dublin, Ireland; [12]Department of General Practice, University College Cork, Cork, Ireland and [13]Department of General Practice, RCSI University of Medicine and Health Sciences, Dublin, Ireland

Abstract

Background. Psychometric methods are used to remove underperforming items and reduce error in existing measures, albeit different approaches can produce different results. This study aimed to determine the implications of applying different psychometric methods for clinical trial outcomes.

Methods. Individual participant data from 15 antidepressant treatment trials from Vivli.org were analyzed. Baseline (pretreatment) and 8-week (range 4–12 weeks) outcome data from the Montgomery-Asberg Depression Rating Scale were subjected to best-practice factor analysis (FA), item response theory (IRT), and network analysis (NA) approaches. Trial outcomes for the original summative scores and psychometric-model scores were assessed using multilevel models. Percentage differences in Cohen's d effect sizes for the original summative and psychometrically modeled scores were the effects of interest.

Results. Each method produced unidimensional models, but the modified scales varied from 7 to 10 items. Treatment effects ($d = 0.072$) were unchanged for IRT (10 items), decreased by 1.3%–2.8% (eight-item abbreviated $d = 0.070$; weighted score $d = 0.071$) for NA, and increased by 11%–12.5% (seven-item abbreviated model $d = 0.081$; weighted score $d = 0.080$) for FA.

Discussion. IRT and NA yielded negligible differences in effect outcomes relative to original trials. FA increased effect sizes and may be the most effective method for identifying the items on which placebo and treatment group outcomes differ.

Introduction

Different psychometric methods and their utility

The development of the questionnaires and rating scales that are used in contemporary psychiatry treatment outcome research usually involves some form of advanced psychometric analysis. Historically, latent variable theory approaches, such as exploratory and confirmatory factor analyses (EFA and CFA, respectively) have predominated in this area of research. These approaches assume that participant scores on outcome measures are a function of latent constructs, and that some of the items on a questionnaire or rating scores may be better indicators of latent constructs than others (De Champlain, 2010). The aim of these techniques is to identify the best set of items and then assess the validity and reliability of the resulting model in measuring the latent construct.

Item response theory (IRT) focuses on how well individual items discriminate along a latent trait. Participants' item responses are measured to determine their 'ability' on the latent trait, with the assumption that higher abilities increase the probability of endorsing respective items (Reise & Waller, 2009). Network analysis (NA) eschews causal latent traits and instead posits that observed items (e.g. symptoms) are mutually causal. Graphical networks of intercorrelated

© The Author(s), 2025. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike licence (http://creativecommons.org/licenses/by-nc-sa/4.0), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the same Creative Commons licence is used to distribute the re-used or adapted article and the original article is properly cited. The written permission of Cambridge University Press must be obtained prior to any commercial use.

CAMBRIDGE
UNIVERSITY PRESS

'nodes' are used to assess the items' relationships (edges) and importance (centrality) within the network (Borsboom, 2017; Borsboom & Cramer, 2013). Each of these methods can be used to identify and remove nonperforming scale items and used to derive weighted outcome scores from resulting models. The weighted scores may be more accurate measures of the construct of interest than are unweighted total scores (Chalmers et al., 2023; Golino et al., 2024; Rosseel et al., 2024).

These types of psychometric methods have played a key role in the development and assessment of the most trusted and commonly adopted outcome measures used in social science research. For example, they have been used to assess validity and reliability of the most commonly used depression rating measures, including the Montgomery-Asberg Depression Rating Scale (MADRS; Quilty, Robinson, Rolland, Fruyt, & Rouillon, 2013), the Beck Depression Inventory-II (Beck, Steer, & Garbin, 1988), the Patient-Reported Outcome Measurement Information System (Nolte, Coon, Hudgens, & Verdam, 2019), and the Hamilton Rating Scale for Depression (HRSD; Hamilton, 1960). Such studies demonstrate the importance of psychometrics in relation to valid and reliable measurement, with the MADRS noted among the optimally performing measures. However, due to their somewhat niche standing and high degree of technical difficulty, the important role they play in good measurement may not be fully recognized by researchers (Sijtsma, 2012), and the lack of a strong connection between psychometrics as a field and substantive psychological research (Wijsen, Borsboom, & Alexandrova, 2022) might mean that more tangible implications of psychometric methods have gone unexplored.

Do psychometrics matter in clinical trial outcomes? Existing evidence

While theoretical arguments of the importance of psychometrics have been made from the perspective of good measurement (Sijtsma, 2012; Wijsen et al., 2022), relatively little research has been conducted to compare these directly, or to explore the tangible implications of these methods, such as those that might affect clinical trial outcomes. One such study used factor analysis (FA) and IRT techniques to conduct a psychometric sensitivity analysis of secondary data collected using the Disability of Arm, Shoulder and Hand (DASH) scale. The aim was to determine the stability of study findings over time. The DASH scale was used to measure upper limb function in 382 women receiving either usual care or a novel exercise program after having undergone surgery for breast cancer. Harrison, Hossain, Bruce, and Rodrigues (2023) noted that the psychometric sensitivity analysis supported the original trial findings (Bruce et al., 2021), in that upper limb ability was higher at 12-month follow-up for the exercise group than for the control group. The implication of the original trial outcomes was that exercise improved upper limb mobility. However, this was contradicted by the psychometric findings, which suggested that exercise only prevented the deterioration seen in the control group. This is an important distinction that sheds light on the true usefulness and efficacy of the exercise program, and the applicability of psychometrics in trials.

Harrison, Hossain, et al. (2023) also examined the implications of using model-based weighted scores by reweighting individual item scores according to their contribution to an IRT model. While scores performed similarly to the original analysis at 12 months, no difference was found between the control and exercise groups at 6 months, contrary to the original study findings. Similar research conducted by Harrison, Plessen, et al. (2023) examined data for the Total or Partial Knee Arthroplasty Trial, which used the Oxford Knee Score to assess 528 patients undergoing total or partial knee replacement and used IRT techniques to reweight item responses. There were statistically significant differences in sum score outcomes for total and partial knee replacement patients, but no statistically significant differences in reweighted scores. A study of RCT and simulated data conducted by Gorter et al. (2016) and Gorter, Fox, Apeldoorn, and Twisk (2016) suggested that analyses using sum scores could be biased to a degree of roughly one standard deviation and that within-person variance tends to be overestimated, while between person variance tends to be underestimated. Although there was no evidence of significant differences in outcome scores between groups, Gorter et al. argued that IRT-weighted scores were a closer reflection of the true scores, as they better accounted for bias. These studies have some limitations. For example, Harrison, Hossain, et al. (2023) and Harrison, Plessen, et al. (2023) used single trial data, which may not be generalizable. Gorter et al. (2016) only assessed mean differences and did not assess potential differences in standardized effect size outcomes, which are superior to simple mean score differences as a measure of the magnitude of observed differences (Sullivan & Feinn, 2012).

More recently, one study examined the effects of different psychometric methods on antidepressant trial outcomes that used the HRSD (Byrne et al., 2025). Data for 6,843 participants in 20 different trials were combined and FA, IRT, and NA were used to obtain optimal abbreviated scales. Original sum scores and abbreviated data were analyzed to identify potential differences in effect size outcomes. While the difference between sum score and IRT outcomes were negligible, FA yielded a 10% *increase* in standardized mean differences between antidepressants and placebo, while NA produced a 15% *decrease* in effects. Interestingly, outcomes using model-derived weighted scores were similar to those of simple sum scores of the abbreviated scale from which they were derived, suggesting weighted scores might not offer additional utility. However, the HRSD-17 is known for its poor psychometric performance (Bagby, Ryder, Schuller, & Marshall, 2004; Broen et al., 2015; Desseilles et al., 2012; Byrne, Doyle, et al., 2024), so further work is required with a measure that has demonstrated better psychometric validity to determine if applying such competing approaches truly matters.

The present study

This study addressed these knowledge gaps and limitations by conducting psychometric analyses of the MADRS pre- and post-treatment using a pooled multi-trial sample. The MADRS is a clinician-rated depression outcome measure consisting of 10 items that assess a range of mood, thought, and neurovegetative symptoms and was introduced in the 1970s to provide an assay of depression symptom severity superior to prevailing measures like the HRSD (Quilty et al., 2013). The MADRS was selected for its previously noted stability and superior performance, when compared to other measures (Carmody et al., 2006; Carneiro, Fernandes, & Moreno, 2015). The aim of this study was to examine the impact of applying advanced psychometric methods on clinical trial effect size outcomes.

Materials and methods

Dataset

Individual participant data (IPD) for the MADRS were obtained from the online clinical trial data repository Vivli.org. Inclusion criteria were participants older than 18 years of age in phase two,

https://doi.org/10.1017/S0033291725101785 Published online by Cambridge University Press

three, or four randomized controlled trials (RCTs) of major or minor depression. Any antidepressant medication was acceptable as the treatment, but only placebo was included as the comparator to ensure outcomes were compared with a uniform control group. The outcome measurement occasion was at 8 weeks after baseline, with a range of 4–12 weeks, as per Cipriani et al. (2017). These inclusion/exclusion criteria and methods were similar to those used previously (Byrne et al., 2025), with the exception that the MADRS was the outcome measure of interest instead of the HRSD. In an additional deviation from this protocol (Doyle et al., 2023), we had also originally intended to analyse data from two separate repositories, but this was not possible due to lack of access to some of the data (see Supplementary Item 1).

Psychometric analysis

FA, IRT, and NA techniques were used to determine an optimal, potentially abbreviated, version of the MADRS according to each approach. Psychometric analyses were conducted using R v4.1.1 in R Studio v1.4.1717 (R Core Team, 2013). FA was conducted as outlined previously (Byrne, Doyle, et al., 2024; Doyle et al., 2023) and involved parallel analysis of a randomly split exploratory group (n = 3,481) to determine dimensionality and factor structure. Items that did not make a sufficient contribution to a latent depression trait were removed according established best practices (Costello & Osborne, 2005), and models were confirmed in relation to several fit indices using a confirmatory group (n = 3,481) (Schermelleh-Engel & Moosbrugger, 2003; Smith & McMillan, 2001). Similarly, IRT involved Mokken analysis of the exploratory group to determine the structure and dimensionality of the data (Crisan, Tendeiro, & Meijer, 2021), with graded response modeling (Chalmers et al., 2023) conducted using the confirmatory group. NA was conducted according to a published protocol (Byrne, Ghoshal, et al., 2024a) and involved exploratory graphical analyses and bootstrapping techniques recommended by Christensen and Golino (2021). The CFA and IRT methods used are outlined in detail in Supplementary Items 2 and 3, respectively. Detailed NA methods are available elsewhere (Byrne, Ghoshal, et al., 2024a).

Weighted scores were then derived from the optimal (abbreviated) models found by each method. CFA 'factor scores' were calculated using the 'lavPredict' function in Lavaan v0.6–9 (Gorter et al., 2015). IRT 'expected scores' were computed using the 'expected test' function in MIRT v1.36.1 (Chalmers et al., 2023), and NA 'net scores' were calculated using the 'net.score' function in EGAnet 1.1.0 (Golino et al., 2024).

Overall, this resulted in the original total scores being compared with abbreviated CFA, IRT, and NA sum scores, as well as weighted scores derived from respective models.

Effect size analysis

Multilevel regressions were used according to best practice methods (Dickenson & Basu, 2005) to determine the effect sizes for the collated data in relation to each of the above outcome scores (Byrne et al., 2025; Doyle et al., 2023). Outcome scores were predicted adjusting for baseline scores, with treatment group as the independent variable and study as the random intercept. Standardized mean differences, calculated as Cohen's d (Cohen, 1988), were obtained for outcomes using the original trial sum scores, as well as each of the abbreviated and weighted scores. These were then compared and percentage differences noted to determine if psychometrically informed effect sizes differed from original trial effects. Multilevel models were also adjusted for potentially moderating demographic variables, including age and sex (Li et al., 2023; Wagner et al., 2020). A detailed description of the effect size analysis plan is available with a published protocol (Doyle et al., 2023).

Results

Sample characteristics

A search of the Vivli.org database found 15 studies (n = 7,009) that met the inclusion criteria (Supplementary Table 1). Of these, 6,962 complete cases were retained for analyses. As the number of cases with missingness was very small (n = 47, 0.6%), sensitivity analysis was not performed and missing data imputation was not considered. Data from each trial were collated into a single analysis file and a variable was created to randomly split participants into exploratory (n = 3,481) and confirmatory (n = 3,481) groups (Doyle et al., 2023) using the *rand()* function in Microsoft Excel. Demographic characteristics, including age, sex and treatment type, can be seen in Table 1.

Psychometric outcomes

The three psychometric methods specified different optimum models, each of which were unidimensional. CFA factor loadings, IRT Loevinger's H coefficients and discrimination parameters, NA centrality parameters and McDonald's Omega reliability coefficients are presented in Table 2. Results for each method are briefly outlined later. More detail on FA and IRT outcomes is available in Supplementary Items 2 and 3, respectively, while NA outcomes are available elsewhere (Byrne, Ghoshal, et al., 2024b).

Factor analysis

EFA found a unidimensional seven-item scale for all outcome models, removing 'Reduced Sleep', 'Reduced Appetite', and 'Suicidal Thoughts'. EFA at baseline initially further removed 'Inner Tension' and 'Concentration Difficulties'. However, this led to configural noninvariance between baseline and outcome models. The outcome model showed optimal performance, so this was retained for baseline. CFA factor loadings for the retained seven items were acceptable at outcome, although 'Inner Tension' was subthreshold at baseline.

IRT modeling

IRT retained all 10 items in a unidimensional model at outcome but initially retained only 'Apparent Sadness' and 'Reported Sadness' at baseline. The outcome model was again examined at baseline to achieve configural invariance. The 10 items performed acceptably at outcome but poorly at baseline, with all items presenting with inadequate H values and relatively poor discrimination coefficients.

Network modeling

Network modeling specified four-community 10-item model at baseline, which bootstrapping found to be unstable. Further analyses indicated a single community eight-item model, which removed 'Pessimistic Thoughts' and 'Suicidal Thoughts.' This network was stable and configurally invariant with outcome models.

Effect size outcomes

Multilevel modeling using all 10 items showed a statistically significant difference between placebo and active treatment groups but presented with a small effect size ($p < 0.001$, $d = 0.072$). The

https://doi.org/10.1017/S0033291725101785 Published online by Cambridge University Press

Table 1. Sample age, sex, and treatment characteristics

		Overall		Exploratory group		Confirmatory group	
		n	%	*n*	%	*n*	%
Age	18–29	1,180	16.9	621	17.8	559	16.1
	30–39	1,542	22.1	732	21.0	810	23.2
	40–49	1,952	28.0	986	28.3	966	27.7
	50–59	1,650	23.7	826	23.7	824	23.7
	60–69	566	8.1	281	8.0	285	8.2
	70+	72	1.2	35	1.1	37	1.1
Sex	Male	2,568	36.9	1,275	36.7	1,293	37.1
	Female	4,394	63.1	2,206	63.3	2,188	62.9
Treatment	Placebo	2,260	32.4	1,132	32.5	1,128	32.4
	Desvenlafaxine	2,250	32.3	1,122	32.3	1,128	32.4
	Duloxetine	146	2.2	77	2.3	69	2.0
	Lu AA21004	1,397	20.0	698	20.0	699	20.0
	TAK–375SL	343	4.9	172	4.9	171	4.9
	Vortioxetine	566	8.2	280	8.0	286	8.3
		6,962		3,481		3,481	

Table 2. Psychometric performance of abbreviated MADRS for baseline and outcome IRT, NA, and CFA models

		Baseline				Outcome			
		CFA	IRT		NA	CFA	IRT		NA
		λ	H	a	nl	λ	H	a	nl
x01	Apparent sadness	0.572	0.183	1.566	0.340	0.897	0.638	4.399	0.440
x02	Reported sadness	0.673	0.252	2.088	0.508	0.911	0.650	4.761	0.475
x03	Inner tension	0.161	0.097	0.403	0.071	0.644	0.513	1.737	0.228
x04	Reduced sleep		0.079	0.378	0.096		0.448	1.259	0.200
x05	Reduced appetite		0.091	0.328	0.127		0.415	1.148	0.169
x06	Concentration Difficulties	0.321	0.142	0.791	0.239	0.697	0.554	1.971	0.287
x07	Lassitude	0.437	0.170	1.055	0.305	0.784	0.605	2.588	0.373
x08	Inability to feel	0.481	0.181	1.205	0.284	0.817	0.608	2.923	0.373
x09	Pessimistic thoughts	0.275	0.143	0.596		0.702	0.556	2.008	
x10	Suicidal thoughts		0.088	0.213			0.447	1.284	
	Total scale		0.135				0.549		
	McDonald's Omega	**0.578**	**0.510**		**0.511**	**0.919**	**0.913**		**0.903**
	Lower ci	0.557	0.487		0.494	0.915	0.908		0.900
	Upper ci	0.600	0.534		0.528	0.923	0.917		0.906
		Placebo outcome				Treatment outcome			
		CFA	IRT		NA	CFA	IRT		NA
		λ	H	a	nl	λ	H	a	nl
x01	Apparent sadness	0.890	0.644	1.566	0.425	0.899	0.632	4.511	0.447
x02	Reported sadness	0.916	0.657	2.088	0.487	0.907	0.645	4.624	0.467
x03	Inner tension	0.663	0.514	0.403	0.232	0.630	0.511	1.673	0.224
x04	Reduced sleep		0.477	0.378	0.212		0.434	1.258	0.196

(*Continued*)

Table 2. (*Continued*)

		Placebo outcome				Treatment outcome			
		CFA	IRT		NA	CFA	IRT		NA
		λ	H	a	nl	λ	H	a	nl
x05	Reduced appetite		0.390	0.328	0.161		0.427	1.193	0.177
x06	Concentration difficulties	0.693	0.537	0.791	0.283	0.695	0.546	1.962	0.290
x07	Lassitude	0.771	0.611	1.055	0.366	0.790	0.599	2.672	0.377
x08	Inability to feel	0.812	0.615	1.205	0.377	0.816	0.603	2.922	0.369
x09	Pessimistic thoughts	0.692	0.544	0.596		0.703	0.559	2.024	
x10	Suicidal thoughts		0.458	0.213			0.437	1.335	
	Total scale		0.552				0.546		
	McDonald’s omega	**0.918**	**0.910**		**0.903**	**0.918**	**0.913**		**0.902**
	Lower ci	0.911	0.902		0.897	0.913	0.907		0.897
	Upper ci	0.925	0.918		0.909	0.923	0.918		0.906

λ, factor loading from CFA; a, discrimination parameter from GRM; H, Loevinger’s coefficient of homogeneity from Mokken; nl, net loading (strength centrality) from NA.

Table 3. Effect size outcomes for multilevel linear modeling of total, abbreviated, and weighted depression scores

		Mean diff.	95% ci	*t*	*p*	Cohen’s *d*	%
Crude	MADRS total	2.240	(1.513, 2.967)	6.048	0.000	0.072	
	CFA abbreviated (7 items)	1.850	(1.315, 2.385)	6.775	0.000	0.081	12.5
	IRT abbreviated	–	–	–	–	–	–
	NA abbreviated (8 items)	1.849	(1.236, 2.462)	5.892	0.000	0.070	−2.8
	CFA factor scores	1.713	(1.213, 2.213)	6.715	0.000	0.080	11.0
	IRT expected scores	2.084	(1.410, 2.758)	6.045	0.000	0.072	0.0
	NA net scores	0.258	(0.174, 0.342)	6.599	0.000	0.071	−1.3
Adjusted	MADRS total	2.249	(1.522, 2.976)	6.059	0.000	0.072	
	CFA abbreviated (7 items)	1.854	(1.319, 2.469)	6.778	0.000	0.081	12.5
	IRT abbreviated	–	–	–	–	–	–
	NA abbreviated (8 items)	1.854	(1.239, 2.469)	5.901	0.000	0.070	−2.8
	CFA factor scores	1.717	(1.217, 2.217)	6.733	0.000	0.080	11.0
	IRT expected scores	2.094	(1.402, 2.786)	6.480	0.000	0.072	0.0
	NA net scores	0.259	(0.175, 0.343)	6.599	0.000	0.071	−1.3

Weighted scores are derived from respective abbreviated models.
%, Percentage change in d from HRSD-17 total; 95% ci, confidence interval for mean difference; d, Cohen’s d; Mean diff., Mean difference in treatment outcomes; p, Significance for mean diff.

FA-informed abbreviated model (seven items) resulted in a 12.5% increase in effect size ($d = 0.081$) and the NA-informed abbreviated model (eight items) saw a 2.8% decrease in effect ($d = 0.071$). IRT retained all 10 items. Weighted scores derived from each model informed similar outcomes: CFA factor scores +11%, NA net scores −1.4% and IRT expected scores yielded no change. Effect size outcomes for multilevel models that adjusted for age and sex reflected the results for crude models (Table 3).

Discussion

Findings

These findings demonstrate the value of applying FA approaches to Patient-reported outcome measures (PROMs) in randomized trials, but little was gained from other approaches. In this large sample of individual participant data from multiple trials, we demonstrated an 11%–12.5% increase in antidepressant effects when applying FA approaches to measurement. IRT saw no change in effects and applying NA and net scores reduced effect sizes by 1.3%–2.8%. However, there was no additional change using model-derived weighted scores over simply abbreviated total scores. All results remained stable when adjusting for age and sex.

However, an alternative and important consideration is whether the FA model may in fact exaggerate the true effect sizes, and thus potentially reduce external validity. Currently, the findings largely support the results of Byrne et al. (2025), with similar magnitudes and directions of effect size change noted when a similar approach investigated changes in effects in the HRSD-17. The present work was conducted using the MADRS, which has been found to outperform the previously used HRSD-17 under psychometric analysis (Carmody et al., 2006; Carneiro et al., 2015). Our findings show the importance of FA for identifying optimal items on which placebo and active treatment outcome scores differ, with subsequent abbreviated models yielding larger effect sizes. Our findings also support those of Byrne et al. (2025) in suggesting that psychometric methods differentially affect antidepressant trial effect size outcomes, with both IRT and NA informing negligible effect size changes. However, with regard to NA outcomes, our findings differ, in that Byrne et al. saw a decrease in effect.

Contrary to the poor psychometric performance of the HRSD-17 noted by Byrne, Doyle, et al. (2024), abbreviated models of the MADRS were found that reflected the latent depression trait and performed well with outcome groups. Similar issues with baseline fit were observed; however, the MADRS still significantly outperformed the HRSD-17 in this regard. Each method indicated different unidimensional models. FA found a 7-item model, removing 'Reduced Sleep', 'Reduced Appetite' and 'Suicidal Thoughts,' and NA retained eight items, with 'Pessimistic Thoughts' and 'Suicidal Thoughts' being removed. IRT retained all 10 items. Reliability analyses reflected psychometric performance, with outcome models showing very good reliability and baseline being suboptimal.

Implications of findings

FA techniques were found to be most influential in moderating effects by removing psychometrically underperforming items and thusly increasing effect size. FA differed from the other methods used by uniquely removing 'Reduced Sleep' and 'Reduced Appetite.' As IRT and NA resulted in negligible differences from original trial outcomes, it can be suggested that the removal of these items was influential in the subsequently increased effects. Although 'Suicidal Thoughts' was also removed by FA, this item being retained by IRT and removed by NA, with each of these methods bearing a negligible impact on effect size analyses, indicated that the presence or absence of a suicide ideation items has little impact on trial outcomes. This is possibly due to this often being an exclusion criterion for such studies. Indeed, 10 of the 15 included studies observed risk of suicide as an exclusion criterion (Studies numbered 1, 9, 10, 11 and 14 in Supplementary Table 1 did not include suicide ideation in exclusion criteria). Furthermore, weighted score outcomes reflecting abbreviated score outcomes suggests that the effect size differences seen after FA are mainly the result of eliminating poorly performing items. Although Gorter, Fox, Apeldoorn, and Twisk (2016) argue the importance of the large mean differences found when using weighted scores, our findings support Byrne et al. (2025) in demonstrating that any notable effect size difference is predicated on the items analyzed (and those removed) and not the weighted scores derived therein. In this regard, FA methods were optimally able to identify items on which placebo and active treatment groups differ, allowing for the removal of nonperforming items, thus informing a measurable percentage increase in effect size compared to original outcomes. As such, these findings suggest that sum score statistics – albeit, those derived from optimal, potentially abbreviated, scales – are sufficient for effect size analyses. This brings into question the utility of deriving and analyzing model-weighted scores. Indeed, this practice has inherent complications, such as factor score indeterminacy (Ferrando & Lorenzo-Seva, 2018), and has been the topic of a long-running debate as to its relative value (Beauducel, Hilger, & Kuhl, 2023; Glass & Maguire, 1966; Rozeboom, 1988; Steiger, 1996).

The psychometric and effect size analyses suggest that removal of 'Reduced Sleep' and Reduced Appetite' caused the FA-informed increase in effect size. The removal of sleep and appetite regulation items was also previously found (Byrne et al., 2025) and may indicate that such items do not discriminate well between placebo and active treatment outcomes. This finding is interesting, as sleep disturbance in particular has been found to be a frequent and salient residual symptom after otherwise successful treatment of major depression (Carney, Freedland, Steinmeyer, & Rich, 2023). However, previous psychometric analyses of the MADRS have raised questions about the performance of these items. For example, a previous study that conducted a factor analysis of the MADRS found that, while both items were retained, Reduced Sleep ($r = 0.57$) and Reduced Appetite ($r = 0.59$) had the lowest single-item correlations with MADRS total scores (Seemüller et al., 2023). Similarly, network analysis of a community-based sample conducted by An et al. (2019) found that these two items presented with the lowest centrality indices of all 10 items. Considering that these items were removed during FA, informing increased effects, and their performance was relatively poor when retained in IRT and NA models, not to mention the previous literature, it stands to reason that although appetite and sleep disturbance may be notable residual symptoms, they are not necessarily an important aspect of patient depression profiles in terms of assessing treatment efficacy. Perhaps unsurprisingly, the NA outcomes observed here reflect previous studies in suggesting that treatment of depressive disorders may be most efficacious when targeting sadness/low mood symptoms (Bringmann, Lemmens, Huibers, Borsboom, & Tuerlinckx, 2015; Maciaszek, Pawlowski, Hadryś, & Misiak, 2023; Park et al., 2021).

It is also noteworthy that 'Suicidal Thoughts' was removed from the FA and NA models, and was one of the least discriminatory items in IRT. This conflicts with previous research that found suicide ideation to be an important symptom in depression profiles (An et al., 2019), as well as an evidence-based narrative that suicide ideation should always be monitored in RCTs of psychotropic substances (Melvin, Gordon, & Freake, 2012; Schatten et al., 2020). As previously argued (Byrne, Ghoshal, et al., 2024b), participants at risk of suicide are typically excluded from antidepressant treatment RCTs. This likely informed the discrepancy between models found here and in Byrne et al. (2025) using RCT samples, and those found by An et al. (2019) using a community-based sample. As such, the FA- and NA-revised MADRS models may only be appropriate for use in RCTs or other types of studies that control for suicide ideation.

This study reflected Byrne et al. (2025) in finding percentage differences between original and psychometrically informed effect size outcomes within the intervals of 10%–15%, and so this could tentatively be proposed as an expected moderating range. Ultimately, it is difficult to determine the implications of the magnitude of this difference. However, such changes could be clinically significant, particularly at larger effect sizes. Research has indicated that the maximum achievable effect when active treatment group patients achieve a 50% symptom reduction over placebo is $d = 1.08$ (Hieronymus et al., 2021). Amending this according to the FA findings, which saw a 12.5% increase in effect, Hieronymus et al.'s maximum symptom reduction-informed effect would be increased to $d = 1.22$. This highlights the potential extrapolation of the percentage change in effect and, as previously suggested (Byrne et al., 2025), could influence treatment efficacy expectations and prescription confidence, as well as trial sample size requirements. However, it should be noted that, while applying FA models in the

https://doi.org/10.1017/S0033291725101785 Published online by Cambridge University Press

future could yield small increased effect sizes which may be important at a population level (rather than individual-level detection of functional benefit), a caveat is that such an abbreviated measure will only capture responses to items measured, thereby masking any potential benefits on symptoms not assessed.

Limitations and future research

A notable limitation of this study was the suboptimal psychometric performance of the MADRS at baseline. The issues encountered modeling baseline data may have reduced the amount of measurement error that could have been controlled and obscured or otherwise limited the potential effect size change that could have been observed (Fumio, 2000). In addition, the effect sizes observed during outcome analyses were much smaller than would typically be expected from antidepressant treatment trials (Cipriani et al., 2017). This makes it more difficult to interpret the implications of the effect size analyses. These issues could be addressed in future research using alternative data, with larger effect sizes. There was also a moderately uneven sex distribution in participants. Considering the increased tendency for women to exhibit depressive symptoms, and to an increased severity (Kokras & Dalla, 2017), findings may be more representative of female populations than male. As NA outcomes presented here conflict with previous findings, future research could also further explore the utility of NA methods in clinical trials. A further limitation is the removal of items that are clinically important, such as disturbed sleep and appetite, which significantly impact on patients and care. Therefore, future trials require better performing items across the full range of depressive symptoms to truly determine the impact of antidepressants and other treatments on each symptom. Additionally, with such small effects, it is probably not possible to determine the benefit of such changes at an individual level. These may, however, be important at a population level, which should be explored in future research. Finally, research should be conducted to further assess the external validity of FA findings and to ensure that these are not in fact simply inflating true effect sizes. In this regard, additional research should systematically assess a broad range of depression outcome measures to determine if outcomes from this study and Byrne, Doyle, et al. (2024) can be replicated.

Conclusion

We demonstrated that that applying FA approaches increased effect sizes in antidepressant trials that used the MADRS, but applying IRT and NA approaches did not. FA methods were most effective in identifying MADRS items where placebo and treatment group outcomes differed, leading to increased effect sizes when compared to original trial outcomes. Weighted scores from FA, IRT, and NA models should not replace traditional total scores from the MADRS, or possibly other scales (Byrne, Doyle et al., 2024). Using simple total scores from abbreviated scales, once nonperforming items are removed, may be a practical alternative for improving sensitivity to group differences. However, other depression scales with stronger psychometric properties may be more clinically relevant and preferable for routine use.

Supplementary Material. The supplementary material for this article can be found at http://doi.org/10.1017/S0033291725101785.

Funding statement. Funded by an Irish Research Council (IRC) 'Collaborative Alliances for Societal Challenges (COALESCE/2021/68)', #LoveIrishResearch.

Competing interests. This publication is based on research using data from data contributors Eli Lilly, GlaxoSmithKline, Pfizer, and Takeda that has been made available through Vivli, Inc. Vivli has not contributed to or approved, and Vivli, Eli Lilly, GlaxoSmithKline, Pfizer, and Takeda are not in any way responsible for, the contents of this publication.

References

An, M. H., Park, S. S., You, S. C., Park, R. W., Park, B., Woo, H. K., Kim, H. K., & Son, S. J. (2019). Depressive symptom network associated with comorbid anxiety in late-life depression. *Frontiers in Psychiatry*, **10**(856):1–10. https://doi.org/10.3389/fpsyt.2019.00856.

Bagby, R. M., Ryder, A. G., Schuller, D. R., & Marshall, M. B. (2004). The Hamilton depression rating scale: Has the gold standard become a lead weight? *American Journal of Psychiatry*, **161**(12), 2163–2177. https://doi.org/10.1176/appi.ajp.161.12.2163.

Beauducel, A., Hilger, N., & Kuhl, T. (2023). The trade-off between factor score determinacy and the preservation of inter-factor correlations. *Educational and Psychological Measurement*, **84**(2), 289–313. https://doi.org/10.1177/00131644231171137.

Beck, A. T., Steer, R. A., & Garbin, M. G. J. (1988). Psychometric properties of the Beck depression inventory twenty-five years of evaluation. *Clinical Psychology Review*, **8**, 77–100. https://doi.org/10.1016/0272-7358(88)90050-5.

Borsboom, D. (2017). A network theory of mental disorders. *World Psychiatry*, **16**(1), 5–13. https://doi.org/10.1002/wps.20375.

Borsboom, D., & Cramer, A. O. J. (2013). Network analysis: An integrative approach to the structure of psychopathology. *Annual Review of Clinical Psychology*, **9**, 91–121. https://doi.org/10.1146/annurev-clinpsy-050212-185608.

Bringmann, L. F., Lemmens, L. H., Huibers, M. J., Borsboom, D., & Tuerlinckx, F. (2015). Revealing the dynamic network structure of the Beck depression inventory-II. *Psychological Medicine*, **45**(4), 747–757. https://doi.org/10.1017/s0033291714001809.

Broen, M. P. G., Moonen, A. J. H., Kuijf, M. L., Dujardin, K., Marsd, L., Richard, I. H., et al. (2015). Factor analysis of the Hamilton depression rating scale in Parkinson's disease. *Parkinsonism & Related Disorders*, **21**(2), 142–146. https://doi.org/10.1016/j.parkreldis.2014.11.016.

Bruce, J., Mazuquin, B., Canaway, A., Hossain, A., Williamson, E., Mistry, P., et al. (2021). Exercise versus usual care after non-reconstructive breast cancer surgery (UK PROSPER): Multicentre randomised controlled trial and economic evaluation. *BMJ*, **375**, e066542. https://doi.org/10.1136/bmj-2021-066542.

Byrne, D., Doyle, F., Brannick, S., Carney, R. M., Cuijpers, P., Dima, A. L., Freedman, K. E., et al. (2024). Evaluating the psychometric structure of the Hamilton rating scale for depression pre- and post-treatment in antidepressant randomised trials: Secondary analysis of 6843 individual participants from 20 trials. *Psychiatry Research*, **339**, 116057. https://doi.org/10.1016/j.psychres.2024.116057.

Byrne D., Ghoshal A., Boland F., Brannick S., Carney R.M., Cuijpers P., Dima A. L., et al. (2024a). Exploring the effects of network analysis on depression trial outcomes: Protocol for secondary analysis of individual participant data [Internet]. *PsyArXiv*. osf.io/preprints/psyarxiv/pr9bv.

Byrne D., Ghosal A., Boland F., Brannick S., Carney R.M., Cuijpers P., Dima A. L., et al. (2024b). An exploratory graphical analysis of the Montgomery-Åsberg Depression Rating scale pre- and post-treatment using pooled antidepressant trial secondary data. *Journal of Affective Disorders*, **368**, 584–590. https://doi.org/10.1016/j.jad.2024.09.087.

Byrne D., Boland F., Brannick S., Carney R.M., Cuijpers P., Dima A., Freedland K.E., et al. (2025). Applying advanced psychometric approaches yields differential randomised trial effect sizes: Secondary analysis of individual participant data from antidepressant studies using the Hamilton Rating Scale for Depression. *Journal of Clinical Epidemiology*, 111762. https://doi.org/10.1016/j.jclinepi.2025.111762.

Carmody, T., Rush, A. J., Bernstein, I., Warden, D., Brannan, S., Burnham, D., Woo, A., et al. (2006). The Montgomery Åsberg and the Hamilton ratings of depression: A comparison of measures. *European Neuropsychopharmacology*, **16**(8), 601–611. https://doi.org/10.1016/j.euroneuro.2006.04.008.

Carneiro, A. M., Fernandes, F., & Moreno, R. A. (2015). Hamilton depression rating scale and Montgomery– Åsberg depression rating scale in depressed and bipolar I patients: Psychometric properties in a Brazilian sample. *Health and Quality of Life Outcomes*, **13**(42), 1–8. https://doi.org/10.1186/s12955-015-0235-3.

Carney, R. M., Freedland, K. E., Steinmeyer, B. C., & Rich, M. W. (2023). Symptoms that remain after depression treatment in patients with coronary heart disease. *Journal of Psychosomatic Research*, **165**, 111–122. https://doi.org/10.1016/j.jpsychores.2022.111122.

Chalmers, P., Pritikin, J., Robitzsch, A., Zoltak, M., Kim, K. H., Falk, C. F., et al. (2023). Multidimensional Item Response Theory. MIRT. https://cran.r-project.org/web/packages/mirt/mirt.pdf.

Christensen, A. P., & Golino, H. (2021). Estimating the stability of psychological dimensions via bootstrap exploratory graph analysis: A Monte Carlo simulation and tutorial. *The Psychiatrist*, **3**(3), 479–500. https://doi.org/10.3390/psych3030032.

Cipriani, A., Furukawa, T. A., Salanti, G., Chaimani, A., Atkinson, L. Z., Ogawa, Y., Leucht, S., et al. (2017). Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder: A systematic review and network meta-analysis. *Lancet*, **391**(10128), 1357–1366. https://doi.org/10.1016/S0140-6736(17)32802-7.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. (2nd ed.) Lawrence Erlbaum Associates.

Costello, A. B., & Osborne, J. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practice, Assessment, Research, and Evolution*, **10**[Art 7], 1–9. https://doi.org/10.7275/jyj1-4868.

Crisan, D. R., Tendeiro, J. N., & Meijer, R. R. (2021). The Crit coefficient in Mokken scale analysis: A simulation study and an application in quality-of-life research. *Quality of Life Research*, **31**, 49–59. https://doi.org/10.1007/s11136-021-02924-z.

De Champlain, A. F. (2010). A primer on classical test theory and item response theory for assessments in medical education. *Medical Education*, **44**, 109–117. https://doi.org/10.1111/j.1365-2923.2009.03425.x.

Desseilles, M., Perroud, N., Guillaume, S., Jaussent, I., Genty, C., Malafosse, A., & Courtet, P. (2012). Is it valid to measure suicidal ideation by depression rating scales? *Journal of Affective Disorders*, **136**, 398–404. https://doi.org/10.1016/j.jad.2011.11.013.

Dickenson, L. M., & Basu, A. (2005). Multilevel modeling and practice-based research. *Annals of Family Medicine*, **3**(1), 52–60. https://doi.org/10.1370/afm.340.

Doyle, F., Byrne, D., Carney, R. M., Cuijpers, P., Dima, A., Freedland, K. E., Guerin, S., et al. (2023). The effects of advanced factor analysis approaches on outcomes in randomised trials for depression: Protocol for secondary analysis of individual participant data. *BJPsych Open*, **9**(5), 1–5. https://doi.org/10.1192/bjo.2023.544.

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.

Fumio, H. (2000). *Econometrics*. Princeton University Press.

Glass, G. V., & Maguire, T. O. (1966). Abuse of factor scores. *American Educational Research Journal*, **3**(4), 297–304. https://doi.org/10.2307/1162038.

Golino H., Christensen C., Moulder R., Garrido L.E., Jamison L., Shi D. (2024). 'EGAnet'. Exploratory Graph Analysis – A framework for estimating the number of dimensions in multivariate data using network psychometrics. https://cran.r-project.org/web/packages/EGAnet/EGAnet.pdf.

Gorter, R., Fox, J. P., Apeldoorn, A., & Twisk, J. (2016). Measurement model choice influenced randomized controlled trial results. *Journal of Clinical Epidemiology*, **79**, 140–149. https://doi.org/10.1016/j.jclinepi.2016.06.011.

Gorter, R., Fox, J. P., Twisk, J. W. R. (2015). Why item response theory should be used for longitudinal questionnaire data analysis in medical research. *BMC Med Res Methodol*, **15**, 55. https://doi.org/10.1186/s12874-015-0050-x.

Hamilton, M. (1960). A rating scale for depression. *Journal of Neurology, Neurosurgery, and Psychiatry*, **23**(1), 56–62. https://doi.org/10.1136/jnnp.23.1.56.

Harrison, C. J., Hossain, A., Bruce, J., & Rodrigues, J. N. (2023). Psychometric sensitivity analyses can identify bias related to measurement properties in trials that use patient-reported outcome measures: A secondary analysis of a clinical trial using the disabilities of the arm, shoulder, and hand questionnaire. *Journal of Clinical Epidemiology*, **163**, 21–28. https://doi.org/10.1016/j.jclinepi.2023.09.008.

Harrison, C. J., Plessen, C. Y., Liegl, G., Rodrigues, J. N., Sabah, S. A., Cook, J. A., Beard, D. J., et al. (2023). Item response theory may account for unequal item weighting and individual-level measurement error in trials that use PROMs: A psychometric sensitivity analysis of the TOPKAT trial. *Journal of Clinical Epidemiology*, **158**, 62–69. https://doi.org/10.1016/j.jclinepi.2023.03.013.

Hieronymus, F., Lisinski, A., Hieronymus, M., Naslund, J., Eriksson, E., & Ostergaard, S. D. (2021). Determining maximal achievable effect sizes of antidepressant therapies in placebo-controlled trials. *Acta Psychiatrica Scandinavica*, **144**(3), 300–309.

Kokras, N., & Dalla, C. (2017). Preclinical sex differences in depression and antidepressant response: Implications for clinical research. *Journal of Neuroscience Research*, **95**(1–2), 731–736. https://doi.org/10.1002/jnr.23861.

Li, S., Zhang, X., Cai, Y., Zheng, L., Pang, H., & Lou, L. (2023). Sex difference in incidence of major depressive disorder: An analysis from the global burden of disease study 2019. *Annals of General Psychiatry*, **22**(53), 1–9. https://doi.org/10.1111/acps.13340.

Maciaszek, J., Pawlowski, T., Hadryś, T., & Misiak, B. (2023). Baseline depressive symptoms as predictors of efficacy and tolerability of the treatment with duloxetine: A network analysis approach. *Frontiers in Psychiatry*, **14**, 1–7. https://doi.org/10.3389/fpsyt.2023.1210289.

Melvin, G. A., Gordon, M. S., & Freake, B. M. (2012). Assessment of suicidal ideation and behavior in clinical trials: Challenges and controversies. *Clinical Investigation*, **2**(3), 265–273. https://doi.org/10.4155/CLI.12.10.

Nolte, S., Coon, C., Hudgens, S., & Verdam, M. G. E. (2019). Psychometric evaluation of the PROMIS® depression item Bank: An illustration of classical test theory methods. *Journal of Patient-Reported Outcomes*, **3**(46), 1–10. https://doi.org/10.1186/s41687-019-0127-0.

Park, S. C., Kin, Y., Kim, K., Woo, Y. S., Kim, J. B., & Jang, E. Y. (2021). Network analysis of the symptoms of depressive disorders over the course of therapy: Changes in centrality measures. *Psychiatry Investigation*, **18**(1), 48–58. https://doi.org/10.30773/pi.2020.0367.

Quilty, L. C., Robinson, J. R., Rolland, J. P., Fruyt, F. D., & Rouillon, F. (2013). The structure of the Montgomery–Åsberg depression rating scale over the course of treatment for depression. *International Journal of Methods in Psychiatry Research*, **22**(3), 175–184. https://doi.org/10.1002/mpr.1388.

R Core Team. (2013). *R:* A language and environment for statistical computing [computer program]. R Foundation for Statistical Computing. https://www.r-project.org/.

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.

Rosseel Y., Jorgensen T.D., Rockwood N., Oberski D., Byrnes J., Vanbrabant L., Savalei L., et al. (2024). 'The lavaan tutorial'. The Lavaan Project. https://lavaan.ugent.be/tutorial/tutorial.pdf.

Rozeboom, W. W. (1988). Factor indeterminacy: The saga continues. *British Journal of Mathematical and Statistical Psychology*, **41**, 209–226.

Schatten, H. T., Guadiano, B. A., Primack, J. M., Arias, S. A., Armey, M. F., Miller, I. W., Epstein-Lubow, G., & Weinstock, L. M. (2020). Monitoring, assessing, and responding to suicide risk in clinical research. *Journal of Abnormal Psychology*, **129**(1), 64–69. https://doi.org/10.1111/j.2044-8317.1988.tb00897.x.

Schermelleh-Engel, K., & Moosbrugger, M. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit

https://doi.org/10.1017/S0033291725101785 Published online by Cambridge University Press

measures. *Methods of Psychological Research Online*, **8**(2), 23–74. https://doi.org/10.23668/psycharchives.12784.

Seemüller, F., Schennach, R., Musil, R., Obermeier, M., Adli, M., Bauer, M., Brieger, P., et al. (2023). A factor analytic comparison of three commonly used depression scales (HAMD, MADRS, BDI) in a large sample of depressed inpatients. *BMC Psychiatry*, **23**(Art 548), 1–12. https://doi.org/10.1186/s12888-023-05038-7.

Sijtsma, K. (2012). Future of psychometrics: Ask what psychometrics can do for psychology. *Psychometrika*, **77**(1), 4–20. https://psycnet.apa.org/doi/10.1007/s11336-011-9242-4.

Smith, T. D., McMillan, B. F. (2001). A primer of model fit indices in structural equation modeling. *Paper presented at Annual meeting of the Southwest Educational Research Association*, New Orleans. https://eric.ed.gov/?id=ED449231.

Steiger, J. H. (1996). Dispelling some myths about factor indeterminacy. *Multivariate Behavioural Research*, **31**(4), 539–550. https://psycnet.apa.org/doi/10.1207/s15327906mbr3104_7.

Sullivan, G. M., & Feinn, R. (2012). Using effect size – Or why P value is not enough. *Journal of Graduate Medical Education*, **4**(3), 279–282. https://doi.org/10.4300/JGME-D-12-00156.1.

Wagner, S., Wollschager, D., Dreimuller, N., Engelmann, J., Herzog, D. P., Roll, S. C., Tadic, A., et al. (2020). Effects of age on depressive symptomatology and response to antidepressant treatment in patients with major depressive disorder aged 18 to 65 years. *Comprehensive Psychiatry*, **99**, 152170. https://doi.org/10.1016/j.comppsych.2020.152170.

Wijsen, L. D., Borsboom, D., & Alexandrova, A. (2022). Values in psychometrics. *Perspectives on Psychological Science*, **17**(3), 788–804. https://doi.org/10.1177/17456916211014183.

https://doi.org/10.1017/S0033291725101785 Published online by Cambridge University Press