human
reproduction
open

ORIGINAL ARTICLE

Reliability and validity of the PHQ-8 in first-time mothers who used assisted reproductive technology

C. Pavlov [1,*], K. Egan[2], and C. Limbers[1]

[1]Department of Psychology and Neuroscience, Baylor University, Waco, TX, USA [2]Peninsula Behavioral Health, Palo Alto, CA, USA

*Correspondence address. Department of Psychology and Neuroscience, Baylor University, One Bear Place #97334, Waco, TX 76798, USA. E-mail: Christina_McCollum1@baylor.edu https://orcid.org/0000-0002-6208-4137

STUDY QUESTION: Is the Patient Health Questionnaire-8 (PHQ-8) a valid and reliable measure of depression in first-time mothers who conceived via ART?

SUMMARY ANSWER: The results from this study provide initial support for the reliability and validity of the PHQ-8 as a measure of depression in mothers who have conceived using ART.

WHAT IS KNOWN ALREADY: Women who achieved a clinical pregnancy using ART experience many stressors and may be at an increased risk of depression. The PHQ-8 is a brief measure designed to detect the presence and severity of depressive symptoms. It has been validated in many populations; however, it has not been validated for use in this population.

STUDY DESIGN, SIZE, DURATION: This is a cross-sectional study of 171 first-time mothers in the USA, recruited through Amazon's Mechanical Turk (MTurk).

PARTICIPANTS/MATERIALS, SETTING, METHODS: The reliability of the PHQ-8 was measured through a Cronbach's alpha, the convergent validity was measured through the correlation between the PHQ-8 and the Generalized Anxiety Disorder-7 (GAD-7) measure of anxiety symptoms, and the structural validity was measured through a Confirmatory Factor Analysis.

MAIN RESULTS AND THE ROLE OF CHANCE: The Cronbach's alpha for the total PHQ-8 was acceptable ($\alpha = 0.922$). The correlation between the PHQ-8 and the GAD-7 was large ($r = 0.88$) indicating good convergent validity. Ultimately, a bifactor model provided the best model fit ($\chi^2(13) = 23.8$, $P = 0.033$; Comparative Fit Index = 0.987; Root Mean Square Error of Approximation = 0.07, Tucker–Lewis Index = 0.972).

LIMITATIONS, REASONS FOR CAUTION: The results are limited by: the predominantly white and well-educated sample, a lack of causation between the use of artificial reproductive technology and depressive symptoms, including mothers with children up to 5 years old, convergent validity being based on associations with a related construct instead of the same construct, lack of test-retest reliability, divergent validity and criterion-related validity, data collected through MTurk, and the fact that the measures used were all self-report and therefore may be prone to bias.

WIDER IMPLICATIONS OF THE FINDINGS: Consistent with previous literature, a bifactor model for the PHQ-8 was supported. As such, when assessing depression in first-time mothers who conceived via ART, using both the PHQ-8 total score and subdomain scores may yield the most valuable information. The results from this study provide preliminary support for the reliability and validity of the PHQ-8 as a measure of depression in first-time mothers who conceived using ART.

STUDY FUNDING/COMPETING INTEREST(S): No specific funding was used for the completion of this study. Throughout the study period and manuscript preparation, the authors were supported by the department funds at Baylor University. The authors declare that they have no conflicts of interest.

TRIAL REGISTRATION NUMBER: N/A.

Key words: PHQ-8 / assisted reproductive technologies / infertility / mothers / depression

WHAT DOES THIS MEAN FOR PATIENTS?

Due to the many stressors that face first-time mothers who conceive using infertility treatments, they may be at risk for depression. While they may be at high risk, little is known about whether our existing measures of depression behave the same way in this population.

This study examines a common measure of depression (the Patient Health Questionnaire-8 or PHQ-8) to see if it accurately and reliably captures depressive symptoms in first-time mothers who conceived with the use of infertility treatments.

After examining the results of the statistical processes, we are able to provide support for the use of the PHQ-8 in this population. Furthermore, we provide support for breaking the PHQ-8 into two subscales to provide valuable information for this population.

Introduction

The International Committee for Monitoring Assisted Reproductive Technologies (ICMART) defines infertility as a disease that is characterized by the failure to establish a clinical pregnancy after 12 months of regular, unprotected sexual intercourse or an impairment of a person's capacity to reproduce either as an individual or with his/her partner (ICMART, 2017). Rates of infertility are rising with one in six couples who want to conceive being diagnosed with infertility (Ravitsky and Kimmins, 2019). ART, which is defined as all interventions that involve the *in vitro* handling of both human oocytes and sperm or of embryos for the purpose of reproduction (ICMART, 2017), increases the likelihood of achieving a clinical pregnancy in couples experiencing infertility. Nonetheless, these treatments are financially costly (Collins, 2001; Katz et al., 2011) and physically and psychologically burdensome (Aimagambetova et al., 2020) especially for females as ART procedures (i.e. daily shots, hormone treatments, egg retrieval, embryo transfer) are largely performed on the woman. Women who have achieved a clinical pregnancy using ART may be at an increased risk for experiencing depression (Ross et al., 2011; Gdańska et al., 2017). This heightened risk for depression and avoidance of negative feelings may continue during the transition to parenthood especially for first-time mothers who may be more likely to idealize parenthood, experience greater concerns about their child's health, and feel less entitled to seek social support when they feel doubts or uncertainty about parenting (Ulrich et al., 2004; Fisher et al., 2005; Gressier et al., 2015). Furthermore, continued infertility and challenges to conceive subsequent children naturally may have a negative impact on the psychological well-being of mothers after conceiving via ART (Hjelmstedt et al., 2004). As such, it is important to have psychometrically sound measures than can assess depressive symptoms in mothers who have conceived using ART, particularly during the transition to parenthood.

The eight-item Patient Health Questionnaire (PHQ-8) is a brief measure designed to detect the presence and severity of depressive symptoms in adults (Kroenke et al., 2002). This measure has demonstrated acceptable internal consistency reliability, test–retest reliability, construct validity, factorial invariance and concurrent validity in Mexican and Central American descent university students residing in the USA, adults from Sweden with Systematic Sclerosis, and Latino/a university students living in the USA (Alpizar et al., 2018a,b; Mattsson et al., 2020). To the best of our knowledge, the psychometric properties of the PHQ-8 have not been previously evaluated in a sample of mothers who conceived using ART. Given increasing rates of infertility (Ravitsky and Kimmins, 2019) and the potential of a greater propensity for depression among first-time mothers who conceived via ART during the transition to parenthood (Ross et al., 2011; Gdańska et al., 2017), the current study sought to evaluate the reliability and validity of the PHQ-8 in first-time mothers of children 5 years old or younger who conceived using ART.

Materials and methods

Participants

The data used in this study were collected as a part of a larger study focused on assessing differences in maternal ratings of child vulnerability between first-time mothers who conceived using ART versus spontaneous conception (Egan et al., 2021). For the current study focused on assessing the psychometric properties of the PHQ-8 in first-time mothers who conceived using ART, only mothers who used ART were included. The sample consisted of 171 first-time mothers. Participants met inclusion criteria if they lived in the USA, were at least 18 years old, a first-time mother of a singleton child 5 years old or younger, endorsed experiencing infertility (which was defined as a failure to attain a clinical pregnancy after 12 months or more of trying to conceive), and reported utilizing a form of ART (i.e. IVF, ICSI, donor egg IVF, gestational carrier IVF, intrauterine embryo implantation, frozen embryo transfer, gamete intrafallopian transfer or zygote intrafallopian transfer) that resulted in the live birth of their child. Mothers of children up to 5 years old were included due to research suggesting that the effects of infertility and ART are far reaching and long lasting (Schmidt, 2010).

Procedures

Participants for this study were recruited during Spring 2018 using Amazon's Mechanical Turk (MTurk). To ensure data quality, the study was advertised as one about parenting and conception methods and a screening survey was administered (Chandler and Shapiro, 2016; Thomas and Clifford, 2017). Once study eligibility was determined from the screening survey, an online consent form was presented to the participant. Participants who provided their online consent to participate in the study were offered the full set of questionnaires including the PHQ-8 and the Generalized Anxiety Disorder-7 (GAD-7).

After completing the survey, each participant was assigned a unique code to verify their participation through Qualtrics and receive compensation through MTurk. Participants received $1.81 for participation in the study and were not allowed to participate more than once.

Ethical approval

The study procedures outlined above were approved by the authors' Institutional Review-Board (IRB) ID#1395596-1.

Measures

Depression

Self-reported depression was measured by the PHQ-8 (Kroenke et al., 2009). The 9-item Patient Health Questionnaire (PHQ-9), from which the PHQ-8 is derived, is a widely used assessment for presence and severity of depressive symptoms (Kroenke et al., 2002). One item included on the PHQ-9 assesses suicide ideation and due to the inability of researchers to adequately respond to de-identified participants reporting suicide ideation, the PHQ-8 was developed with this item excluded (Kroenke et al., 2009). Exclusion of this item does not it influence the sensitivity of the measure in detecting major depression (Kroenke et al., 2009). The PHQ-8 has been validated as a diagnostic tool and measure of depressive symptoms in clinical settings and large surveys in populations (Kroenke and Spitzer, 2002; Kroenke et al., 2009).

Participants were asked to reflect on their past 2 weeks and respond to 8 items on a 4-point Likert-type scale ranging from 'not at all' (0) to 'nearly every day' (3). Example items include: 'little interest or pleasure in doing things', 'feeling tired or having little energy' and 'feeling down, depressed or hopeless'. Item responses were subsequently totaled and measured on a scale ranging from 0 to 24. Scores were interpreted as follows: 0–4 (minimal/no depression), 5–9 (minimal depression), 10–14 (moderate depression), 15–19 (moderately severe depression), 20–24 (severe depression) (Kroenke and Spitzer, 2002; Kroenke et al., 2009, 2010).

Anxiety

Self-reported anxiety was measured by the GAD-7 (Spitzer et al., 2006). This 7-item assessment has been validated to measure a unidimensional factor of general anxiety disorder in the general population (Löwe et al., 2008; Naeinian et al., 2011). Participants were asked to reflect upon the last 2 weeks and answer, on a Likert-type scale ranging from 'not at all' (0) to 'nearly every day' (3), how often they had been bothered by the following problems: (i) feeling nervous, anxious, or on edge, (ii) not being able to stope or control worrying, (iii) worrying too much about different things, (iv) trouble relaxing, (v) being so restless that it is hard to sit still, (vi) becoming easily annoyed or irritable and (vii) feeling afraid, as if something awful might happen. Item responses were subsequently totaled and measured on a scale ranging from 0 to 21. Scores were interpreted as follows: 0–4 (minimal anxiety), 5–9 (mild anxiety), 10–14 (moderate anxiety), 15–21 (severe anxiety) (Spitzer et al., 2006).

The GAD-7 has been used to support convergent validity in previous validation studies of measures of depression (Löwe et al., 2008). Specifically, the GAD-7 and the PHQ-9 have been shown to be strongly correlated (Quon et al., 2015; Sawaya et al., 2016; Peters et al., 2021; Sequeira et al., 2021). Anxiety has been previously shown to be significantly correlated with depression in women who have conceived using ART (Huang, et al., 2019). Therefore, in order to assess convergent validity through the measurement of an associated construct, the GAD-7 was included in this study.

Demographic questionnaire

Participants were asked to respond to a questionnaire assessing the following information: maternal age, age of first child, income, education, employment, maternal age at child's birth, incidence of miscarriage, whether or not their child was born prematurely, marital status, number of ART treatments, cause of infertility, whether or not their insurance covered their ART treatments and what type of ART they used to conceive their child.

Statistical analysis

IBM Statistical Package for the Social Sciences (SPSS) Version 26 was used for this project. Statistical significance was determined by a P-value <0.05. An examination of skewness was conducted for the PHQ-8 and the GAD-7. Due to the nature of MTurk, there were no missing data and no data were excluded.

Floor and ceiling effects

An examination of floor and ceiling effects was conducted for the PHQ-8. Floor and ceiling effects were determined by examining whether or not greater than 15% of participants received either the lowest (floor) or highest (ceiling) possible score (McHorney and Tarlov, 1995; Terwee et al., 2007). Results with floor or ceiling effects indicate potentially poor content validity (Terwee et al., 2007).

Internal consistency reliability

Internal consistency reliability was assessed by examining Cronbach's alphas for the PHQ-8. An acceptable range for Cronbach's alpha is a value of 0.70 or higher (Nunnally, 1978). Inter-item correlations and the modified Cronbach's alpha associated with the deletion of each item were also examined. Correlations were considered small $r \leq 0.1$, medium $r \geq 0.3$ and large $r \geq 0.5$ (Cohen, 1988).

Convergent validity

Convergent validity may be defined as the magnitude of the zero-order correlation between two closely related measures (Carlson and Herdman, 2012). To assess convergent validity, Pearson correlations were examined between the PHQ-8 and the GAD-7 (Spitzer et al., 2006). Correlations were considered small $r \leq 0.1$, medium $r \geq 0.3$ and large $r \geq 0.5$ (Cohen, 1988).

Structural validity

Structural validity of the PHQ-8 was assessed through a Confirmatory Factor Analysis (CFA) using R version 3.6.1. There is evidence from a Monte Carlo Simulation to suggest that for a multiple-factor model with 6–8 indicators, the minimum sample size needed is 100 participants (Wolf et al., 2013). Therefore, the sample included in this study of 171 was sufficient. The R code for the CFA can be found in the Supplementary Information. The development literature for the PHQ-8 suggests that the scale is measuring one factor (Kroenke et al., 2009). However, previous literature has suggested that in some populations, a two-factor model may present a more accurate model fit (Mattsson et al., 2020). Furthermore, there is evidence from work done with both the PHQ-8 and PHQ-9 that suggests a bifactor model, which estimated model fit based on a general factor of depression as well as two latent variables, is the superior model (Doi et al., 2018; Dong et al., 2019; Fischer et al., 2021). Therefore, this study examined a single factor, a two-factor, and a bifactor model. For both the two-factor model

and bifactor model, the two latent variables specified were cognitive/affective aspects of depression (items 1, 2, 6 and 7) and somatic aspects of depression (items 3, 4, 5 and 8) (Mattsson *et al.*, 2020).

Within the CFA, the chi-square statistic (Hu and Bentler, 1999) was examined along with other measures of model fit including the Root Mean Squared Error of Approximation (RMSEA) Comparative Fit Index (CFI) and Tucker–Lewis index (TLI). In order to achieve excellent model fit, RMSEA values must be equal to or less than 0.06 and in order to achieve acceptable model fit, RMSEA values should be <0.08 (Browne and Cudeck, 1992; Hu and Bentler, 1999). In order to achieve excellent model fit, CFI and TLI values must be equal to or greater than 0.95 and in order to achieve acceptable fit, CFI and TLI values should range between 0.90 and 0.95 (Mulaik *et al.*, 1989; Bentler, 1990; Hu and Bentler, 1995).

Results

Demographic data

Socio-demographic characteristics for this sample are provided in Table I. Most participants were around 30 years old (SD = 4.65,

Table I Demographic variables for the sample.

Characteristic	N or mean	% or SD	Range
Age	30.37	4.65	22–46
Age of eldest child	1.92	1.30	0–5
Age at birth of first child	28.02	4.63	20–44
Marital status			
Married	148	86.5%	–
Divorced	2	1.2%	–
Single	20	11.7%	–
Separated	1	0.6%	–
Cause of Infertility			
Male-factor	33	19.3%	–
Female-factor	97	56.7%	–
Both	10	5.8%	–
No known cause	31	18.1%	–
Previous instances of miscarriage			
0	86	50.3%	–
1	53	31.0%	–
2	25	14.6%	–
3	6	3.5%	–
4	1	0.6%	–
Was your child born prematurely?			
Yes	40	23.4%	–
No	131	76.36%	–
Did insurance cover your ART treatment cycles?			
No	34	19.9%	–
Yes, partially	83	48.5%	–
Yes, fully	54	31.6%	–

Table I Continued

Characteristic	N or mean	% or SD	Range
Highest level of education			
Some high school	2	1.2%	–
High school degree	9	5.3%	–
Some college	20	11.7%	–
Trade/technical school	1	0.6%	–
Associate degree	15	8.8%	–
Bachelor's degree	89	52.0%	–
Master's degree	28	16.4%	–
Doctorate	7	4.1%	–
Race/ethnicity			
White	130	76%	–
Hispanic/Latino	13	7.6%	–
Black/African American	12	7.0%	–
Asian	12	7.0%	–
American Indian/Alaska Native	1	0.6%	–
Missing	3	1.8%	–
Income			
Under $25 000	12	7.0%	–
$25 000 to $34 999	28	16.4%	–
$35 000 to $49 999	31	18.1%	–
$50 000 to $74 999	44	25.7%	–
$75 000 to $99 999	32	18.7%	–
$100 000 to $149 999	17	9.9%	–
Over $150 000	7	4.1%	–
Employment			
Employed full-time	108	63.2%	–
Employed part-time	39	22.8%	–
Unemployed, looking	2	1.2%	–
Unemployed, not looking	2	1.2%	–
Homemaker	18	10.5%	–
Retired	1	0.6%	–
Disabled, unable to work	1	0.6%	–

N = 171.

Range = 22–46) and were about 28 years old at the time of the birth of their first child (SD = 4.63, Range = 20–44). The majority of participants were married (86.5%), had at least a 4-year degree (72.5%), were employed (86%), and had children over the age of 18 months (65.5%). A large portion of participants also reported an income at or above $35 000 (58.5%), were White (76%), experienced infertility caused by a female factor (56.7%), had experienced one or more miscarriages (49.7%), did not have their child prematurely (76.4%) and received at least some financial help with infertility treatments from insurance (80.1%). The most common form of ART used was IVF with 73.3% of participants reporting having used IVF at some point throughout their infertility treatments. About 49% of women in this sample had undergone one to three cycles of ART. In this sample, 36.4% of mothers reported moderate to severe depressive symptoms.

Table II Interitem correlation matrix and Item Reliability Statistics (PHQ-8).

Item	1	2	3	4	5	6	7	8	Cronbach's alpha if item deleted
1. PHQ-8_1	–								0.910
2. PHQ-8_2	0.61	–							0.912
3. PHQ-8_3	0.58	0.51	–						0.914
4. PHQ-8_4	0.53	0.47	0.58	–					0.918
5. PHQ-8_5	0.64	0.60	0.66	0.58	–				0.910
6. PHQ-8_6	0.61	0.70	0.54	0.54	0.61	–			0.911
7. PHQ-8_7	0.61	0.64	0.62	0.55	0.63	0.55	–		0.911
8. PHQ-8_8	0.71	0.69	0.55	0.56	0.59	0.65	0.63	–	0.909

N = 171; PHQ-8, Patient Health Questionnaire-8.

Floor and ceiling effects

Floor effects on the PHQ-8 occurred in 15.8% of participants; 0.6% of participants reported a ceiling effect on the PHQ-8 total score. For the cognitive/affective subscale of the PHQ-8, floor effects occurred in 31.6% of the population. Ceiling effects occurred in 0.6% of participants on the cognitive/affective subscale of the PHQ-8. For the somatic subscale of the PHQ-8, floor effects occurred in 18.1% of participants. Ceiling effects occurred in 0.6% of the sample on the somatic subscale of the PHQ-8.

Internal reliability

Cronbach's alpha for the PHQ-8 total score in this sample was within the acceptable range ($\alpha = 0.922$). The Cronbach's alphas for the cognitive/affective and the somatic subscales were also within the acceptable range ($\alpha = 0.867$, $\alpha = 0.850$, respectively). Table II presents the inter-item correlations for each of the eight items in the scale along with the Cronbach's alpha if that item were to be deleted. All of the items correlated highly with each other (r's ranged from 0.47–0.71). The Cronbach's alphas would decrease with the deletion of any item in the scale.

Convergent validity

In support of convergent validity, correlation between the PHQ-8 and the GAD-7 was in the large range (GAD-7 $r = 0.88$, $P < 0.001$). Similarly, the cognitive/affective and the somatic subscales of the PHQ-8 correlated strongly with measures of maternal anxiety (cognitive/affective: GAD-7 $r = 0.873$, $P < 0.001$; somatic: GAD-7 $r = 0.811$, $P < 0.001$).

Structural validity

The skewness of each item was assessed and found to fall within the normal distribution.

Fit indices for the CFAs can be found in Table III. Item loadings can be found in Table IV. The CFA testing a one-factor model used maximum-likelihood estimators. The one-factor model demonstrated adequate to excellent fit on most indices, $\chi^2(20) = 52.83$, $P < 0.001$; CFI = 0.961; RMSEA = 0.098; TLI = 0.945. All items loaded significantly onto the latent factor ($\beta > 0.82$).

Table III Fit indices by model.

Model	χ^2	df	CFI	RMSEA	TLI
One-factor Model	52.83	20	0.961	0.098	0.945
Two-factor Model	52.83	19	0.96	0.102	0.941
Bifactor Model	23.8	13	0.987	0.07	0.972

CFI, Comparative Fit Index; RMSEA, Root Mean Square Error Approximation; TLI, Tucker–Lewis Index.

The CFA testing a two-factor model also used maximum-likelihood estimators. While most of the indices for the two-factor model also demonstrated adequate to excellent ($\chi^2(19) = 52.83$, $P < 0.001$; CFI = 0.960; RMSEA = 0.102, TLI = 0.941), the one-factor model demonstrated a superior fit to the two-factor model. All items for the two-factor model loaded significantly onto their designated latent variable (cognitive/affective $\beta > 0.964$; somatic $\beta > 0.899$). The two latent variables had a moderate covariance of (CoV = 0.539).

The CFA testing a bifactor model also used maximum-likelihood estimators. The bifactor model demonstrated adequate to excellent fit on all indices, $\chi^2(13) = 23.8$, $P = 0.033$; CFI = 0.987; RMSEA = 0.07, TLI = 0.972. Not all items loaded significantly onto their designated latent variable (Cognitive/affective: item 7 $\beta = 3.59$, $P = 0.133$; Somatic: item 5 $\beta = 0.367$, $P = 0.271$), but all loaded significantly on the general variable (Depression $\beta > 0.871$, $P < 0.001$). Overall, the bifactor model demonstrated a superior fit to the one-factor and two-factor models.

Discussion

The purpose of the present study was to evaluate the psychometric properties of the PHQ-8 in mothers who conceived using ART. Women who have conceived via ART are at an increased risk for experiencing emotional distress (Aimagambetova et al., 2020). Consistent with previous research (Drosdzol and Skrzypulec, 2009;

Table IV PHQ-8 item loadings by model.

Indicator	One-factor model	Two-factor model		Bifactor model		
		Cog./Aff.	Somatic	Cog./Aff.	Somatic	g
Item 1	1.000	1.000		1.000		1.000
Item 2	1.030	1.030		1.322		1.037
Item 3	0.911		1.000		1.000	0.984
Item 4	0.819		0.899		1.018	0.871
Item 5	0.981		1.077		0.367	1.008
Item 6	1.024	1.024		0.574		1.026
Item 7	0.964	0.964		0.359		0.961
Item 8	0.996		1.094		−1.243	0.891

Item 1: Little interest or pleasure in doing things; **Item 2**: Feeling down, depressed or hopeless; **Item 3**: Trouble falling or staying asleep, or sleeping too much; **Item 4**: Feeling tired or having little energy; **Item 5**: Poor appetite or overeating; **Item 6**: Feeling bad about yourself-or that you are a failure or have let yourself or your family down; **Item 7**: Trouble concentrating on things such as reading the newspaper or watching television; **Item 8**: Moving or speaking so slowly that other people could have noticed-or the opposite-being so fidgety or restless that you have been moving around a lot more than usual. PHQ-8, Patient Health Questionnaire-8; Cog./Aff., Cognitive/Affective; g, effect size Hedges-g.

Ross *et al.*, 2011), in the current sample, 36.4% of mothers reported moderate to severe depressive symptoms. Since maternal depression can be detrimental to both the mother and the child (Cox *et al.*, 1987), it is critical to have psychometrically sound measures that assess depression in mothers who have conceived using ART.

In this population, the PHQ-8 demonstrated good internal consistency reliability. The Cronbach's alphas for the PHQ-8 total score and subdomain scores far exceeded the alpha value of 0.70 recommended for comparing patient scores. Furthermore, the PHQ-8 total score and subdomain scores were highly correlated with measures of maternal anxiety indicating strong convergent validity. There were no ceiling effects for the PHQ-8 total score or subdomain scores in the present study. The PHQ-8 total score and subdomain scores did demonstrate some floor effects. There is evidence that floor effects may be more common in measures that assess symptoms of depression (Tomitaka, *et al.*, 2017; Shin *et al.*, 2020). This may be particularly true when assessing depression in a sample like ours that has a heightened risk for experiencing depression. A recent study conducted among Swedish patients with systemic sclerosis found no floor effects on the PHQ-8. Future studies are needed to evaluate floor effects of the PHQ-8 in other samples of mothers who have conceived via ART, including mothers of older children and adolescents, to determine if there are floor effects that may interfere with the ability of the PHQ-8 to differentiate between individuals who are experiencing high levels of depression.

The CFA analysis of the PHQ-8 revealed that a one-factor model was a better fit than a two-factor model. This result is consistent with previous literature testing single factor and two-factor models for the PHQ-8 (Alpizar *et al.*, 2018a, b). The bifactor model demonstrated the best overall fit in our sample. As such, when assessing depression in first-time mothers who conceived via ART, using both the PHQ-8 total score and subdomain scores may yield the most valuable information. Furthermore, the PHQ-8 total score and subdomain scores were highly correlated with scores on the GAD-7 which assess maternal anxiety symptoms.

Since depression and anxiety are often comorbid (Spitzer *et al.*, 2006) this indicates strong convergent validity ($r = 0.88$, $P < 0.001$). Taken as a whole, data from this study provide preliminary support for utilization of the PHQ-8 as a measure of depression in first-time mothers who have conceived using ART.

This study had a number of limitations. Given that the sample was predominantly white and well-educated, the present findings may not generalize to more diverse mothers who conceived using ART. We were not able to establish in our study that maternal depressive symptoms were a result of the ART experience; it would have been beneficial to include in the study a measure of stressful life events in order to control for other situations that may have been affecting maternal adjustment. Furthermore, our sample was comprised of first-time mothers who had a child 5 years old or younger and it is possible that, given this wide age range of children, mothers in our sample experienced different types of stressors that impacted their psychological well-being. It should be noted that we computed *post hoc* a Pearson correlation between maternal PHQ-8 scores and the age of child; this correlation ($r = -0.14$) was small and not statistically significant ($P = 0.066$), indicating child age may have had a small impact on maternal ratings of their depressive symptoms. Additionally, our assessment of convergent validity was based on a measure of a theoretically related construct, not by measuring the same construct. We also were not able to assess test-retest reliability, divergent validity, and criterion-related validity of the PHQ-8. Another limitation of this study was collecting data via Mechanical Turk. While Mechanical Turk has been shown to yield quality data (Kees *et al.*, 2017) and participants in our study answered screening questions, there was no way for us to objectively determine maternal utilization of ART to conceive. Finally, our study relied entirely on maternal self-reports which may be prone to bias. Future research should focus on assessing test–retest reliability, divergent validity, and criterion-related validity of the PHQ-8 in mothers who have conceived via ART and there would be merit

in reproducing this study in an in-person setting rather than online.

In conclusion, the results from this study provide preliminary support for the reliability and validity of the PHQ-8 as a measure of depression in first-time mothers who conceived using ART.

Supplementary data

Supplementary data are available at *Human Reproduction Open* online.

Data availability

The data underlying this article cannot be shared publicly due to the outlined agreement with the IRB at the author's institution. The data will be shared on reasonable request to the corresponding author.

Authors' roles

C.P., K.E. and C.L.: (i) substantial contributions to conception and design, or acquisition of data, or analysis and interpretation of data, (ii) drafting the article or revising it critically for important intellectual content, (iii) final approval of the version to be published and (iv) agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Funding

Conflict of interest

The authors declare that they have no conflicts of interest.

References

Aimagambetova G, Issanov A, Terzic S, Bapayeva G, Ukybassova T, Baikoshkarova S, Aldiyarova A, Shauyen F, Terzic M. The effect of psychological distress on IVF outcomes: reality or speculations? *PLoS One* 2020;**15**:e0242024.

Alpizar D, Laganá L, Plunkett SW, French BF. Evaluating the eight-item Patient Health Questionnaire's psychometric properties with Mexican and Central American descent university students. *Psychol Assess* 2018;**30**:719–728.

Alpizar D, Plunkett SW, Whaling K. Reliability and validity of the 8-item Patient Health Questionnaire for measuring depressive symptoms of Latino emerging adults. *J Latina/o Psychol* 2018;**6**:115–130.

Bentler PM. Comparative fit indexes in structural models. *Psychol Bull* 1990;**107**:238–246.

Browne MW, Cudeck R. Alternative ways of assessing model fit. *Sociol Methods Res* 1992;**21**:230–258.

Carlson KD, Herdman AO. Understanding the impact of convergent validity on research results. *Organ Res Methods* 2012;**15**:17–32.

Chandler J, Shapiro D. Conducting clinical research using crowd-sourced convenience samples. *Annu Rev Clin Psychol* 2016;**12**: 53–81.

Cohen J. Set correlation and contingency tables. *Appl Psychol Meas* 1988;**12**:425–434.

Collins J. Cost-effectiveness of in vitro fertilization. *Semin Reprod Med* 2001;**19**:279–289.

Cox AD, Puckering C, Pound A, Mills M. The impact of maternal depression in young children. *Journal of Child Psychology and Psychiatry* 1987;**28**:917–928.

Doi S, Ito M, Takebayashi Y, Muramatsu K, Horikoshi M. Factorial validity and invariance of the Patient Health Questionnaire (PHQ)-9 among clinical and non-clinical populations. *PLoS One* 2018;**13**: e0199235.

Dong L, Skolarus L, Morgenstern L, Lisabeth L. Abstract WMP89: Examining the Constructs of the Patient Health Questionnaire (PHQ-8) in the Stroke Population. *Stroke* 2019;**50**:AWMP89.

Drosdzol A, Skrzypulec V. Depression and anxiety among Polish infertile couples—an evaluative prevalence study. *J Psychosom Obstet Gynaecol* 2009;**30**:11–20.

Egan K, Summers E, Limbers C. Perceptions of child vulnerability in first-time mothers who conceived using assisted reproductive technology. *J Reprod Infant Psychol* 2021;1–11.

Evans-Hoeker EA, Eisenberg E, Diamond MP, Legro RS, Alvero R, Coutifaris C, Casson PR, Christman GM, Hansen KR, Zhang H et al.; Reproductive Medicine Network. Major depression, antidepressant use, and male and female fertility. *Fertil Steril* 2018;**109**: 879–887.

Fischer F, Levis B, Falk C, Sun Y, Ioannidis J, Cuijpers P, Shrier I, Benedetti A, Thombs BD; Depression Screening Data (DEPRESSD) PHQ Collaboration. Comparison of different scoring methods based on latent variable models of the PHQ-9: an individual participant data meta-analysis. *Psychol Med* 2021;1–12.

Fisher JR, Hammarberg K, Baker HG. Assisted conception is a risk factor for postnatal mood disturbance and early parenting difficulties. *Fertil Steril* 2005;**84**:426–430.

Gdańska P, Drozdowicz-Jastrzębska E, Grzechocińska B, Radziwon-Zaleska M, Węgrzyn P, Wielgoś M. Anxiety and depression in women undergoing infertility treatment. *Ginekol Pol* 2017;**88**: 109–112.

Gressier F, Letranchant A, Cazas O, Sutter-Dallay AL, Falissard B, Hardy P. Post-partum depressive symptoms and medically assisted conception: a systematic review and meta-analysis. *Hum Reprod* 2015;**30**:2575–2586.

Hjelmstedt A, Widström AM, Wramsby H, Collins A. Emotional adaptation following successful in vitro fertilization. *Fertil Steril* 2004; **81**:1254–1264.

Hu LT, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Struct Equ Model* 1999;**6**:1–55.

Hu LT, Bentler PM. Evaluating model fit. In RH Hoyle (ed). *Structural Equation Modeling: Concepts, Issues and Application*. Thousand Oaks, CA: Sage, 1995, 77–99.

Huang MZ, Kao CH, Lin KC, Hwang JL, Puthussery S, Gau ML. Psychological health of women who have conceived using assisted reproductive technology in Taiwan: findings from a longitudinal study. *BMC Womens Health* 2019;**19**:1–11.

ICMART (2017). https://www.icmartivf.org/glossary/a-d/#A (14 March 2022, date last accessed).

Katz P, Showstack J, Smith JF, Nachtigall RD, Millstein SG, Wing H, Eisenberg ML, Pasch LA, Croughan MS, Adler N. Costs of infertility treatment: results from an 18-month prospective cohort study. *Fertil Steril* 2011;**95**:915–921.

Kees J, Berry C, Burton S, Sheehan K. An analysis of data quality: professional panels, student subject pools, and Amazon's Mechanical Turk. *J Advert* 2017;**46**:141–155.

Kroenke K, Spitzer RL. The PHQ-9: a new depression diagnostic and severity measure. *Psychiatr Ann* 2002;**32**:509–515.

Kroenke K, Spitzer RL, Williams JB, Löwe B. The patient health questionnaire somatic, anxiety, and depressive symptom scales: a systematic review. *Gen Hosp Psychiatry* 2010;**32**:345–359.

Kroenke K, Strine TW, Spitzer RL, Williams JB, Berry JT, Mokdad AH. The PHQ-8 as a measure of current depression in the general population. *J Affect Disord* 2009;**114**:163–173.

Lanzi RG, Bert SC, Jacobs BK; Centers for the Prevention of Child Neglect. Depression among a sample of first-time adolescent and adult mothers. *J Child Adolesc Psychiatr Nurs* 2009;**22**:194–202.

Löwe B, Decker O, Müller S, Brähler E, Schellberg D, Herzog W, Herzberg PY. Validation and standardization of the Generalized Anxiety Disorder Screener (GAD-7) in the general population. *Med Care* 2008;**46**:266–274.

Mattsson M, Sandqvist G, Hesselstrand R, Nordin A, Boström C. Validity and reliability of the Patient Health Questionnaire-8 in Swedish for individuals with systemic sclerosis. *Rheumatol Int* 2020; **40**:1675–1687.

McHorney CA, Tarlov AR. Individual-patient monitoring in clinical practice: are available health status surveys adequate? *Qual Life Res* 1995;**4**:293–307.

Mulaik SA, James LR, Van Alstine J, Bennett N, Lind S, Stilwell CD. Evaluation of goodness-of-fit indices for structural equation models. *Psychol Bull* 1989;**105**:430–445.

Naeinian MR, Shairi MR, Sharifi M, Hadian M. To study reliability and validity for a brief measure for assessing Generalized Anxiety Disorder (GAD-7). *Arch Intern Med* 2011;**166**:1092–1097.

Nunnally JC. Psychometric Theory. 2nd edn, 1978. MC Grew-Hill, New York..

Peters L, Peters A, Andreopoulos E, Pollock N, Pande RL, Mochari-Greenberger H. Comparison of DASS-21, PHQ-8, and GAD-7 in a virtual behavioral health care setting. *Heliyon* 2021;**7**:e06473.

Quon BS, Bentham WD, Unutzer J, Chan YF, Goss CH, Aitken ML. Prevalence of symptoms of depression and anxiety in adults with cystic fibrosis based on the PHQ-9 and GAD-7 screening questionnaires. *Psychosomatics* 2015;**56**:345–353.

Ravitsky V, Kimmins S. The forgotten men: rising rates of male infertility urgently require new approaches for its prevention, diagnosis and treatment. *Biol Reprod* 2019;**101**:872–874.

Ross LE, McQueen K, Vigod S, Dennis CL. Risk for postpartum depression associated with assisted reproductive technologies and multiple births: a systematic review. *Hum Reprod Update* 2011;**17**: 96–106.

Sawaya H, Atoui M, Hamadeh A, Zeinoun P, Nahas Z. Adaptation and initial validation of the Patient Health Questionnaire–9 (PHQ-9) and the Generalized Anxiety Disorder–7 Questionnaire (GAD-7) in an Arabic speaking Lebanese psychiatric outpatient sample. *Psychiatry Res* 2016;**239**:245–252.

Schmidt L. Psychosocial consequences of infertility and treatment. In: Carrell DT, Peterson CM (eds) Reproductive Endocrinology and Infertility. 2010. New York: Springer, pp. 93–100.

Sequeira SL, Morrow KE, Silk JS, Kolko DJ, Pilkonis PA, Lindhiem O. National norms and correlates of the PHQ-8 and GAD-7 in parents of school-age children. *J Child Fam Stud* 2021;**30**: 2303–2312.

Shin C, Ko YH, An H, Yoon HK, Han C. Normative data and psychometric properties of the Patient Health Questionnaire-9 in a nationally representative Korean population. *BMC Psychiatry* 2020; **20**:1–10.

Spitzer RL, Kroenke K, Williams JB, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. *Arch Intern Med* 2006;**166**:1092–1097.

Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, Bouter LM, de Vet HC. Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol* 2007;**60**:34–42.

Thomas KA, Clifford S. Validity and Mechanical Turk: An assessment of exclusion methods and interactive experiments. *Comput Hum Behav* 2017;**77**:184–197.

Tomitaka S, Kawasaki Y, Ide K, Akutagawa M, Yamada H, Yutaka O, Furukawa TA. Item response patterns on the patient health questionnaire-8 in a nationally representative sample of US adults. *Front Psychiatry* 2017;**8**:251.

Ulrich D, Gagel DE, Hemmerling A, Pastor VS, Kentenich H. Couples becoming parents: something special after IVF? *J Psychosom Obstet Gynaecol* 2004;**25**:99–113.

Wolf EJ, Harrington KM, Clark SL, Miller MW. Sample size requirements for structural equation models: an evaluation of power, bias, and solution propriety. *Educ Psychol Meas* 2013;**73**:913–934.