INTERNATIONAL JOURNAL OF SOCIAL SCIENCES BULLETIN
ISSN: (E) 3007-1917 (P) 3007-1909

Volume 4, Issue 3, 2026

IMPLICIT BIAS IN CLINICAL TRAINING: DEVELOPMENT AND VALIDATION OF THE FORMAN SCALE (FIBS)

Aiza Humayun[*1], Prof. Dr. Sarah Shahed[2]

[*1,2]*Department of Psychology, Forman Christian College University, Lahore, Pakistan*

[*1]aiza.humayun@icloud.com

[*1]0009-0008-7354-3176

Corresponding Author: *
Aiza Humayun

DOI: https://doi.org/10.5281/zenodo.21771853

Received	Accepted	Published
17 January 2026	02 March 2026	16 March 2026

ABSTRACT

The current study developed and validated the Forman Implicit Bias Scale for trainee healthcare professionals in the Pakistani cultural context. Based on the Attribution Theory, the study consisted of four phases: Item generation, exploratory factor analysis, establishing psychometric properties, and confirmatory factor analysis. There were eight samples in the study, labeled A to H, each representing a different step. The final scale was a four-factor, reliable, and valid instrument for assessing implicit bias in trainee healthcare professionals during patient interactions. Cut-off scores were also derived to categorize trainee healthcare professionals into low- and high-bias groups. Limitations of the study included an extremely small sample size for confirmatory factor analysis. The scale is primarily based on trainee healthcare professionals, so its use with trained healthcare professionals remains unknown. Despite certain limitations, this scale enables the measurement and identification of implicit bias in healthcare settings.

Keywords: *Implicit bias; Trainee healthcare professionals; Healthcare; Scale development; Exploratory Factor Analysis.*

INTRODUCTION

In a world where the need for healthcare professionals is increasing, it is even more crucial to ensure that these people follow ethical standards. At the core of it, the doctor-patient relationship sets the base of the therapeutic alliance that is formed in any treatment-seeking setting. The four primary elements essential to this relationship, which form the basis of the therapeutic alliance, are mutual knowledge, trust, loyalty, and regard (Ridd et al., 2009). However, when any of these foundational qualities are violated, the therapeutic outcome is affected, too. The strength of the relationship depends on both practitioner- and patient-related factors.

Among the systemic factors is the urgent care setting (Chipidza, Wallwork, & Stern, 2015). Practitioners see multiple patients daily and rely on preconceived notions to fit as many patients as possible into their limited time. Certain group assumptions are automatically activated when individuals from that group seek treatment, especially when the healthcare provider is under pressure, busy, or distracted (Yzerbyt, Rogier, & Fiske, 1998). This hinders the development of a productive therapeutic alliance, especially when the healthcare provider is unaware of the biases they may be unconsciously holding. In psychology, bias is defined as the negative evaluation of one group and its members relative to another (Blair, Steiner & Havranek, 2011). Implicit attitudes were first defined by Greenwald and Banaji (1995) as "introspectively unidentified traces of past experiences that mediate favorable or unfavorable feeling, thought, or action toward social objects." Implicit bias occurs when a link is created between a group or individual's attribute and a negative evaluation or another category attribute (FitzGerald & Hurst, 2017). Social categories may include gender, age, race, ethnicity, socioeconomic status, religion, and mental health status, among others. It has three main characteristics that inform a working definition of the construct. This bias is hidden, occurs unconsciously, and is activated through automatic mental associations.

Essentially, the concept of implicit bias originated in the theory of implicit social cognition in social psychology. When Greenwald and Banaji (1995) introduced the concept of implicit attitudes, the focus shifted from attention to implicit memory. A key feature of implicit memory is that it centers on an individual's performance, making it directly observable and quantifiable, thereby forming an operational definition. This became the foundation of the Implicit Association Test (IAT), an indirect measure of implicit attitudes. However, a major limitation of the IAT is that it does not indicate whether an individual prefers a category, has an aversion to the opposite category, or has a mixture of both.

Although many theories have been proposed, the current study uses attribution theory as the theoretical basis for the newly developed scale. Kelley (1967) proposed the fundamental attribution error, in which people tend to attribute behavior to dispositional factors (personality traits) rather than to situational factors that may have influenced it. In group contexts, this can perpetuate stereotypes. For instance, people often attribute poor performance by an out-group member to innate qualities, which can reinforce bias. Causal schemata, the belief systems individuals hold about cause-and-effect relationships, also play a role. Culture, stereotypes, and unconscious biases heavily influence these schemata. Attribution theory may be used to propose that attributional patterns can influence implicit biases.

In this study, implicit biases are defined as the unconscious negative attitudes and stereotypes that trainee healthcare professionals hold toward patients based on patient characteristics and demographics, such as age, language, gender, socioeconomic status, educational background, mental health status, and weight. These hidden biases affect their verbal behavior (communication style, medical decision-making, and management strategies) and nonverbal behavior (such as eye contact, facial expression, and body language) without conscious awareness. We include psychologists, psychiatrists, and medical doctors who are currently in training when discussing the term "trainee healthcare professionals."

Given the absence of a scale to measure implicit bias among trainee healthcare providers in the Pakistani context, and because the construct affects the patient-practitioner relationship and patient care, it became necessary to measure and report implicit bias in this context using the Forman Implicit Bias Scale (FIBS).

LITERATURE REVIEW

Implicit bias has gained attention over the years in the healthcare industry due to its impact on the patient population. To rectify the adverse outcomes of implicit bias, multiple measures have been constructed for practitioners to use and, consequently, keep in check to improve their relationship with their clients. The current section provides literature on the development of measures for implicit bias, the usage of these measures in research, and their implications for the general population who seek treatment.

When discussing measures of implicit bias, we broadly define them into two categories: Implicit and explicit measures. Explicit measures of implicit bias are self-reported, and the individuals are expected to answer a series of questions related to the negative evaluation and category attribute. However, some researchers believe these measures have elements of social desirability that do not accurately reflect implicit bias. Thus, they developed other procedures.

One of the most famous and frequently used measures of implicit attitudes is the Implicit Association Test (IAT) developed by Greenwald

IEER Institute for Excellence in Education & Research

INTERNATIONAL JOURNAL OF
SOCIAL SCIENCES BULLETIN
ISSN: (E) 3007-1917 (P) 3007-1909
Volume 4, Issue 3, 2026

and Banaji (1995). It aims to measure automatic evaluations that are the foundation of implicit attitudes. The IAT design evaluates the association between target-concept discrimination and an attribute dimension (Greenwald, McGhee, & Schwartz, 1998). There are a series of steps in this procedure. Initially, target-concept discrimination is introduced where the individual must distinguish any concept, such as recognizable names, for example, European American or African American. They must sort the concept into two groups using either responses on the left-hand or right-hand. Next, another categorization task is introduced, focusing on words that are seen as pleasant or unpleasant. This aims to establish the attribute dimension. Once these two tasks are done separately, both categories are assimilated with the first concept, and words appear individually for classification. After this, the individual sees that the responses for categorizing the first concept would reverse, but the word-categorization task doesn't change. Lastly, the two tasks are combined again. In conclusion, it is seen that if the participants associate the concept differently with pleasant or unpleasant words, it will mean an implicit attitude towards the categories. While the IAT is useful in revealing implicit attitudes biased towards specific social categories, some research has shown that quick responses are often made for stimuli that the individual is already familiar with, leading to shorter response latencies (Cvencek et al., 2010). This indicates that latency is not primarily related to the implicit attitude but is attributable to exposure. Additionally, it is an online measure with restricted access to the population in countries such as Pakistan. For most of the population who cannot read English or use computers, the applicability of IAT in the Pakistani context is void.

Another measure of implicit cognition is the Implicit Relational Assessment Procedure (IRAP) developed by Barnes-Holmes et al. (2006). It is based on the Relational Frame Theory (RFT), which proposes that higher-cognitive functioning consists of relational acts (Barnes-Holmes et al., 2006). It says that instead of associations, stimulus relations are essential in comprehending psychological experiences in individuals. It is a computer-based task where participants must answer quickly, which could either align with or contrast with their existing verbal relations. This task assumes that individuals would respond more efficiently to stimuli that mimic their present belief systems than stimuli that do not. IRAP is considered a direct measure of implicit social cognitions. However, IRAP has received much criticism. Firstly, the procedure's reliability is questioned as it overestimates positive implicit attitudes. Moreover, the way that the tasks are phrased has a significant impact on the results. This has also posed a deficit with previous research, such as frequent conflicts with other implicit measures such as IAT. For example, there are reports of neutral attitudes towards categories often believed to be harmful, such as fat people (O'Shea, Watson, & Brown, 2015). Thus, the true essence of implicit bias is not reported.

Among the assessment measures of Implicit Bias, another instrument was developed by Gonzalez et al. (2021) known as the Attitudes Toward Implicit Bias Instrument (ATIBI). It assesses medical students' attitudes to identify and manage their implicit bias. The Delphi technique and factor analysis were conducted, and 5 factors emerged: valuing implicit bias recognition and management training, self-acceptance of bias, self-awareness, bias recognition, and acceptance of bias in clinical care. The limitations of the present study were many, such as the fact that the data was collected only from one institution, lowering generalization. Additionally, since this instrument is an explicit measure, it is more prone to social desirability and may not reflect the actual belief system.

Another self-report measure that is commonly used is the Modern Racism Scale (MRS) developed by McConahay et al. (1986). It measures how much individuals have modern racism beliefs which are typically subtle forms of racism present in people who believe that there is no inequality in the world thus they depict as if they support equality whereas inwardly, they still have those biases against certain groups. The scale has a Likert-based rating system from 1 to 5 and has good internal reliability. However, there has been evidence reporting criticism of the MRS. Firstly,

IEER
Institute for Excellence in Education & Research

there is doubt whether the MRS measures a distinct type of racism or whether it is similar to old-fashioned racism. Additionally, it focuses on the primary problem of discrimination against dark-skinned people and does not account for discrimination faced by other cultural and racial groups. The wording of the MRS is also such that it is difficult to negate the role of social desirability in the results.

We include psychologists, psychiatrists, and medical doctors when discussing healthcare professionals. There have been years and years of research on how implicit bias in these healthcare professionals can affect the outcomes of the patients and how they ultimately go through their disease.

Lawrence et al. (2021) did a systematic review and meta-analysis of 41 studies where the weight-biased attitudes of healthcare professionals were the sole focus. The studies assessed the impact of this bias on patient-provider communication and how care is provided. Results showed that the healthcare professionals exhibited implicit and explicit weight bias towards patients. The healthcare outcomes were also measured, and the results showed low-quality outcomes.

Another study by Abbott et al. (2023) evaluated 82 healthcare professionals by inviting them to a webinar and asking them to use the BiasProof mobile test based on the IAT. The aim was to explore the implicit bias of these professionals who were specializing in obesity and see what their perceptions were about non-weight-focused approaches. Results showed that more than half of the healthcare professionals had an implicit weight bias against people who were living with obesity.

Bruce and Weinraub (2023) investigated the prevalence of implicit gender bias, which impacted the clinical decision-making of psychologists. A cross-sectional design was used to present case vignettes to 180 psychologists, and the gender pronouns were manipulated to assess biases in diagnosing personality disorders. The results of the study showed that in women, borderline personality disorder (BPD) was more likely to be diagnosed, and in men, antisocial personality disorder. Thus, implicit gender bias was proved in the research. However, the sample was relatively limited, and there was no use for real-world interactions between the psychologist and the clients; instead, more reliance was placed on the vignettes. Additionally, only some relevant demographic details of the psychologists were studied, an essential factor affecting clinical decision-making.

Yasmeen, Nazar, and Perveen (2023) studied the implicit bias of age, religion, and skin tone among 100 nurses in one hospital in Pakistan using the Implicit Assessment Tool. Results showed that only 2% preferred dark skin, only 1% preferred Judaism, and 30% preferred young people to older adults. Overall, 96% showed a higher preference for Islam than Christianity. It was concluded that nurses strongly prefer their religion, lighter skin tone, and younger people.

METHODOLOGY

DeVellis's (2017) classical scale construction model was used in the present study. The study comprised four phases. Phase I focused on item generation; Phase II was Exploratory Factor Analysis; Phase III established psychometric properties; and Phase IV was Confirmatory Factor Analysis. Each phase included multiple steps. The data for the entire study are available at https://osf.io/zkaph/overview?view_only=57a1728909d84c9d981ac89a5152db79.Phase I: Item Generation

Methods

The objectives of phase I included conducting a comprehensive literature review and focus group discussions, generating an initial item pool, establishing content validity through expert validation, and ultimately conducting a pilot test of the initial item pool. Deductive item generation drew items from pre-existing scales, whereas inductive item generation relied on focus group discussions. The total item pool contained 203 items in English. A 5-point Likert scale, ranging from 1 = Strongly Disagree to 5 = Strongly Agree, was chosen for its familiarity, ability to detect subtleties, and reliability in attitude measurement (Fowler, 2009; DeVellis, 2017).

Items were refined to ensure methodological accuracy and cultural relevance to Pakistani

IEER Institute for Excellence in Education & Research

culture. An expert review preceded the assessment of face and content validity.

Face Validity. The experts received a comprehensive package that included the 143-item pool, a description of implicit bias, and the rating criteria. Each item was evaluated on relevance, appropriateness, clarity, and social desirability using a 5-point Likert scale.

Content Validity. A panel of six Subject Matter Experts (SMEs) was chosen, all of whom had at least five years of practical experience. They received a similar package to the face validity group. The Item-Content Validity Index (I-CVI) was calculated for all items.

Pilot Study. The purpose of the pilot study was to evaluate the practical usage of the scale. Participants completed the scale and were asked to provide feedback on any items that were unclear in wording or difficult to attempt. The tools given to participants were the sociodemographic questionnaire, the Forman Implicit Bias Scale, the Biased Attitudes Scale, and the Big-Five Inventory (BFI).
The final item pool had 58 items.

Participants

Sample A was the theoretical models and items extracted from preexisting measures of bias.
Sample B consisted of four different focus groups conducted for inductive item generation. This sample had four subsets: B1 to B4. B1 included patients with personal or direct family experience with a healthcare provider, B2 involved trainee psychologists, B3 included trainee medical doctors, and B4 comprised subject matter experts (SMEs) who were experienced professionals from various clinical fields, including psychology, psychiatry, general medicine, and pediatrics.
Sample C for face validity comprised a panel of six experts, including a clinical psychologist and five trainee clinical psychologists.
Sample D was a panel of six SMEs for content validity. They were selected using purposive sampling because they had at least five years of proficiency in clinical psychology, psychiatry, or medical education.
Sample E was the pilot study, consisting of 50 trainee healthcare professionals selected through convenience sampling. It was also used for test-retest reliability. They were taken from academic institutions and hospitals across two disciplines, psychology and general medicine.

Results

The internal review reduced the number of items from 203 to 143 to minimize survey fatigue and reduce fake answers. It improved the content-to-length ratio of the FIBS without sacrificing the theoretical base. This resulted in a 70-item version of the FIBS that was used for the pilot study, as the content validity assessment was ongoing at the time.
Content validity was assessed using the I-CVI. A cutoff of 0.69 was deemed suitable according to Lynn's (1986) recommendation. I-CVI ratings were calculated in Microsoft Excel version 2024 (see data availability statement). After analysis, items with I-CVI scores of 0.69 or higher were retained, and the rest were removed. A few items overlapped with those identified for removal during the face validity assessment. After accounting for this overlap, 13 items were removed from the 70-item version of the FIBS. One item, item 91, had a rating of 0.67 but was retained at the researcher's discretion because it was considered to have significant theoretical relevance.
The pilot study was conducted on Sample E. After item reduction and post-piloting, the final version of the scale consisted of 58 items, which were used in Phase II of the study.

Phase II: Exploratory Factor Analysis
Methods

Exploratory Factor Analysis (EFA) was used to identify the underlying structure of the newly developed scale. Data screening was conducted before EFA. Three items with extreme outliers, which persisted despite many efforts, were removed. All EFA assumptions were also checked. Factorability was assessed through correlation analysis using IBM Statistical Package for the

IEER Institute for Excellence in Education & Research

Social Sciences (SPSS) version 26.0. Two tests, the Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy and Bartlett's Test of Sphericity, were also conducted.

Principal Components Analysis (PCA) was chosen for data reduction as the aim was to condense the data into a few components with high variance. The Kaiser's Criterion Eigenvalues were also checked to see if they were higher than 1. An analysis of the scree plot was also done. Then, Direct Oblimin rotation was used as the study was exploratory in nature.

Participants

Sample F was for EFA, having five times as many participants as the final item pool. The required sample size was 275, given 55 items. They were chosen through purposive sampling from academic institutions and hospitals.

Results

The demographic characteristics of the sample F used in this phase have been presented below.

Table 1

Demographic Characteristics of the Participants in Sample F (N = 275)

	M	SD	f	%
Age	24.66	2.78		
Gender				
Male			51	18.5
Female			221	80.4
Prefer not to say			3	1.1
Religion				
Islam			268	97.5
Christianity			7	2.5
Other			0	0
Ethnicity				
Punjabi			231	84
Pathan/Pakhtoon			8	2.9
Kashmiri			6	2.2
Balochi			6	2.2
Gilgiti			3	1.1
Sindhi			3	1.1
Urdu-speaking			5	1.8
Siraiki			4	1.5
Muhajir			2	.7
Rajput			1	.4
Other			6	2.2
Occupation				
Trainee Clinical Psychologist			150	54.5
House Officer			80	29.1
Postgraduate Resident			20	7.3
Nurse Trainee			20	7.3
Medical Officer			5	1.8
Education				
Bachelor's			111	40.4
Postgraduate			161	58.5
Doctorate (PhD)			3	1.1
Socioeconomic Status				

IEER
Institute for Excellence in Education & Research

Low	2	.7
Lower-Middle	10	3.6
Middle	147	53.5
Upper-Middle	102	37.1
Upper	14	5.1

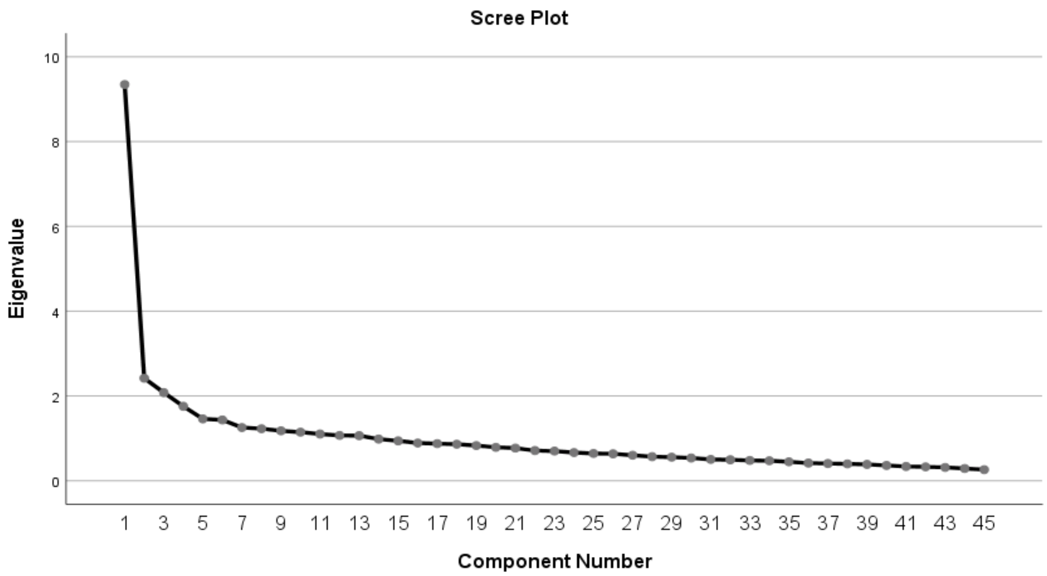
Note: M = mean, SD = standard deviation, f = frequency, and % = percentage

A preliminary analysis using item-total correlations and the correlation matrix was done before EFA. A few items were removed that had low corrected item-total correlations and very weak correlations with other items. The final items for EFA were 45.

The value of Kaiser-Mayer-Olkin (KMO) Measure of Sampling Adequacy and Bartlett's Test of Sphericity was appropriate for factor analysis (Kaiser, 1974). Bartlett's test of sphericity was significant, $\chi^2(990) = 3511.60$, $p < .001$, indicating that the correlation matrix supported the use of factor analysis (Bartlett, 1951).

Principal Component Analysis was performed. The scree plot analysis was done to identify factors that were to be retained.

Figure 1. Scree Plot Explaining the Factor Structure of FIBS.



As suggested by DeVellis (2017), it is recommended to conduct a parallel analysis in IBM SPSS version 26.0 to determine the number of factors to retain (Hayton, Allen, & Scarpello, 2004). This was done to decide how many factors and, consequently, items to extract.

Table 2
Identification of the Number of Retained Factors Using Parallel Analysis

	Number of Factors	Raw Data Eigenvalue	Simulated Mean Eigenvalue	Eigenvalue Percentile
Factor Analyses	1	9.35	1.87	1.98
	2	2.42	1.79	1.84
	3	2.08	1.71	1.77
	4	1.76	1.64	1.70
	5	1.46	1.59	1.63

6	1.43	1.54	1.58
7	1.26	1.49	1.53
8	1.23	1.45	1.49
9	1.18	1.41	1.44
10	1.15	1.37	1.40
11	1.10	1.33	1.37
12	1.07	1.29	1.32
13	1.06	1.26	1.29

Theoretically and statistically, the study found that a four-factor structure was most suitable for the scale. It explained nearly 35% of the total variation. This structure balanced simplicity and stability, as higher-factor models explained more variance but yielded conceptually separate domains with few items in each. It also provided a strong foundation for the confirmatory factor analysis to be conducted later.

Direct Oblimin is the most widely used rotation method in the social sciences and assumes the exploratory nature of a new scale; thus, it was employed in this study. Items were retained based on the following criteria:

- Factor Loadings ≥ 0.30 on a primary factor
- Minimal cross-loadings
- Conceptual coherence
- Absence of redundancy

Consistent with usual procedure, each retained component had at least three well-loaded items to ensure stability and interpretability (Costello & Osborne, 2005). All retained factors had eigenvalues exceeding 1. After analyzing and selecting items with loadings (> .30), the final factor structure was established. The final scale consisted of 4 factors and 33 items.

Table 3
Eigen Values and Variance Explained by Factors of FIBS

Factors	Eigen Values	% of Variance	Cumulative %
Factor 1	9.35	20.8	20.8
Factor 2	2.42	5.38	26.1
Factor 3	2.08	4.62	30.8
Factor 4	1.76	3.90	34.7

Final Factor Structure

Factor 1 – Dismissive Behavior of Practitioners. The items included in this factor are 4, 16, 20, 22, 23, 27, 30, 40, 42, 45, 47, 50, and 54, totaling 13 items. This factor is directly related to practitioners' dismissive behavior toward patients.

Factor 2 – Diagnostic Framing. A total of nine items were retained: 12, 14, 17, 18, 19, 25, 36, 46, and 58. The factor focuses on the diagnosis aspect during healthcare interactions.

Factor 3 – Differential Care Behavior. Five items were retained in this factor: 26, 31, 33, 34, and 38. This component elaborates on the noticeable practitioner behaviors observed during interactions with patients, including tone of voice, eye contact, explanations of the disease/disorder, and observable empathy.

Factor 4 – Devaluation of Patient Credibility. A total of six items were retained: 35, 41, 43, 44, 51, and 52. Concerns about judging a patient's honesty, skill, or insight are part of this factor, especially when the patient is from a minority or marginalized group.

These item numbers are from the original 58-item FIBS. The items of each subscale in the finalized scale are mentioned in the supplementary material.

Inter-Scale Correlation

Pearson Product-Moment Correlation Analysis was conducted for each FIBS subscale and the total score among trainee healthcare professionals. The results showed that the Diagnostic Framing (DF) subscale had a significant, positive, and moderate correlation with the Dismissive Behavior of Practitioners (DBP) subscale, the Differential Care Behavior (DCB) subscale, and the Devaluation of Patient Credibility (DPC) subscale. The DCB subscale had a significant, positive, and strong correlation with the DBP and DPC subscales. The DPC subscale had a significant, positive, and strong correlation with the DBP subscale. All subscales showed a significant, positive, and strong correlation with the overall FIBS. Therefore, the FIBS demonstrated good inter-scale correlation.

Table 4
Inter-Scale Correlation of FIBS

	DBP	DF	DCB	DPC
Dismissive Behavior of Practitioners				
Diagnostic Framing	.45***			
Differential Care Behavior	.54***	.37***		
Devaluation of Patient Credibility	.63***	.45***	.50***	
Full Scale	.83***	.70***	.78***	.82***

Note: DBP = Dismissive Behavior of Practitioners; DF = Diagnostic Framing; DCB = Differential Care Behavior; DPC = Devaluation of Patient Credibility. *** p <.000 level.

Phase III: Establishing Psychometric Properties
Methods
In this phase, construct validity, criterion validity, and reliability of the scale were assessed. Reliability was checked through internal consistency and test-retest. Receiver Operating Characteristic (ROC) analysis was used to establish cut-off scores.
For construct validity, convergent and discriminant validity were established. Convergent validity was measured using the simplification bias subscale from the Biased Attitude Scale (Watts et al., 2020). Discriminant validity was checked using the conscientiousness subscale from the Big Five Inventory (BFI).
The criterion validity variable was the participants' occupational roles. Research has shown that different healthcare positions can be linked to how patients are contacted, how decisions are made, and other factors that can affect implicit attitudes (Chapman et al., 2013; van Ryn & Fu, 2003). It was also observed that the type and years of clinical experience affected the formation and expression of implicit biases among healthcare providers (FitzGerald & Hurst, 2017).
Cronbach's alpha coefficient was computed to assess internal consistency and for test-retest reliability, Pearson's Correlation, and Intra-Class Correlation Coefficient (ICC). The ROC curve helped identify usable cut-off scores to categorize the trainees into two levels of implicit bias (e.g., low and high).

Participants
Sample G, again selected through purposive sampling, comprised 50 trainee healthcare professionals. They were clinical psychology trainees, house officers, and post-graduate residents. They all had at least four months of clinical training. This sample was used for content validity and reliability.
Sample H comprised 400 trainee healthcare professionals. These were psychology trainees, house officers, post-graduate residents, nurse trainees, and medical officers. They had a similar four-month experience to Sample G. Sample H was used to establish criterion validity and cut-off scores through ROC curve analysis.

Results
Construct Validity
Pearson Product-Moment Correlation Analysis was done to identify convergent validity between the newly developed FIBS subscales and the Biased Attitudes (BAS) subscale of Simplification Bias. The results showed a significant, positive, and moderate relationship between the BAS

simplification subscale and the FIBS Differential Care Behavior subscale. This showed moderate convergent validity between the subscale and the established BAS measure.

Table 5
Pearson Product-Moment Correlation between FIBS and BAS Subscale for Convergent Validity (N=50)

	1	2	3	4	5	6
1. Simplification Bias Subscale	-					
2. Dismissive Behavior of Practitioners	-.06	-				
3. Diagnostic Framing	.05	.73***	-			
4. Differential Care Behavior	.40**	.52***	.48***	-		
5. Devaluation of Patient Credibility	.10	.70***	.60***	.57***	-	
6. Full Scale	.15	.87***	.83***	.77***	.87***	-

Note: ****p*<.000; ***p*<.001

No significant correlations were observed for discriminant validity between the Conscientiousness subscale from the BFI and FIBS (see data availability statement for details).

Criterion Validity The criterion variable was the occupation of the participants. All assumptions of one-way ANOVA were fulfilled. The sample was divided into five groups. There was a statistically significant difference at the $p < .05$ level in FIBS scores for the five occupation groups [$F(4, 395) = 3.12$, $p = 0.015$]. The effect size, calculated using eta squared, was 0.031, which was small to borderline medium. Post hoc comparisons showed that there was a statistically significant difference between postgraduate residents (PGRs) and nurse trainees at the $p < 0.01$ level. Nurse trainees ($M = 3.34$) scored significantly higher than the PGRs ($M = 2.85$) on the FIBS scale (Mean difference = -.0487, $p = .005$, Bonferroni adjusted).

Table 6
Mean, Standard Deviation, and One-Way Analysis of Variance of FIBS Scores Across Occupational Groups. (N = 400).

	Occupation											
	TCP		HO		PGR		NT		MO			
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*F (4,395)*	η^2
FIBS	3.02	.52	3.00	.56	2.85	.54	3.34	.50	2.94	.53	3.12*	.024

Note: M = Mean; SD = Standard Deviation; L = Low; LM = Lower-Middle; Mi = Middle; UM = Upper-Middle; H = High. TCP = Trainee Clinical Psychologist; HO = House Officer; PGR = Postgraduate Resident; NT = Nurse Trainee; MO = Medical Officer. F = Anova statistic; η = eta squared.

Reliability

Internal consistency of the scale was checked using Cronbach's alpha, which was .90. According to the results, the Cronbach alpha value was between .90 and .99, indicating excellent internal consistency.

Test-retest reliability analysis was performed on sample E (N = 50) to evaluate the temporal stability of the FIBS. After at least two weeks, the same scale version was given again. For the analysis, Pearson's Correlation and the Intra-Class Correlation Coefficient (ICC) were computed for Time 1 and Time 2 (see the data availability statement for further details). ICC estimates and their 95% confidence intervals were calculated based on absolute agreement and a 2-way mixed-effects model. The results showed that the ICC ranged from .75 to .90, indicating good reliability, and was significant at $p < .000$, reflecting good test-retest reliability for the new scale.

Cut-off Scores using Receiver Operating Characteristic (ROC) Curve Analysis

To establish cut-off scores, a percentile method was used based on the FIBS total score. It helped categorize bias as high or low. The "high bias" was defined as the overall FIBS score being within the top 25th percentile, and "low" if outside that range. The state variable on SPSS was this grouping variable. To avoid circularity, the adjusted total score for each subscale was calculated by adding the mean scores of the other three subscales. To create the binary "high bias" group, the 75^{th} percentile was computed for each adjusted total (top 25% = 1, others = 0). ROC was then conducted using this mean as the test variable and the binary group as the state variable. Area Under the Curve (AUC) was used to assess the diagnostic accuracy.

Table 7
Summary of ROC Analysis for Subscales of FIBS

	AUC	SE	*p*	95% CI	
				LL	UL
DBP	.90	.06	.004**	.77	1.0
DF	.88	.06	.006**	.77	.99
DCB	.69	.14	.17	.42	.96
DPC	.91	.05	.003**	.82	1.0

Note: DBP = Dismissive Behavior of Practitioners; DF = Diagnostic Framing; DCB = Differential Care Behavior; DPC = Devaluation of Patient Credibility; AUC = Area Under the Curve; SE = Standard Error; p = significance value; CI = Confidence Interval; LL = Lower Limit; UL = Upper Limit.

Values over 0.70 were acceptable, above 0.80 good, and above 0.90 exceptional. Youden's Index was also calculated to determine the optimal cut-off score for each subscale. To maintain domain-specific diagnostic value, the well-known Depression, Anxiety, and Stress Scale (DASS), for instance, similarly reports and interprets distinct cut-off scores for each subscale rather than a total composite (Lovibond & Lovibond, 1995). Results showed that the values for three subscales were significant. The Area Under the Curve (AUC) for subscales 1 and 4 falls within the excellent range for differentiating between conditions. The AUC for subscale 2 was in the good ability category. The AUC for subscale 3 was in the poor category. The following table shows the optimal thresholds, i.e., the cut-off scores for the subscales, based on Youden's Index.

Table 8
Cut-off Scores for Each Subscale on FIBS Based on the Youden Index

	Optimal Threshold	Youden Index
DBP	3.65	.69
DF	3.06	.69
DCB	4.30	.44
DPC	3.58	.71

Note: DBP = Dismissive Behavior of Practitioners; DF = Diagnostic Framing; DCB = Differential Care Behavior; DPC = Devaluation of Patient Credibility.

Phase IV: Confirmatory Factor Analysis
Methods

Confirmatory Factor Analysis (CFA) was employed to verify the internal factor structure of FIBS. The four factors derived from EFA, consisting of the final 33 items, were used for confirmation. The model fit indices included a combination of absolute, incremental, and parsimonious fit indices. The model was developed using IBM SPSS AMOS software version 26.

Participants

The convenience sample was 125 trainee healthcare professionals from academic institutions. The sample was below the typical threshold required for CFA.

Results

Once the psychometric properties were evaluated, CFA was done to validate the factor structure of FIBS.

Figure 2. First-Order Model Fit of CFA

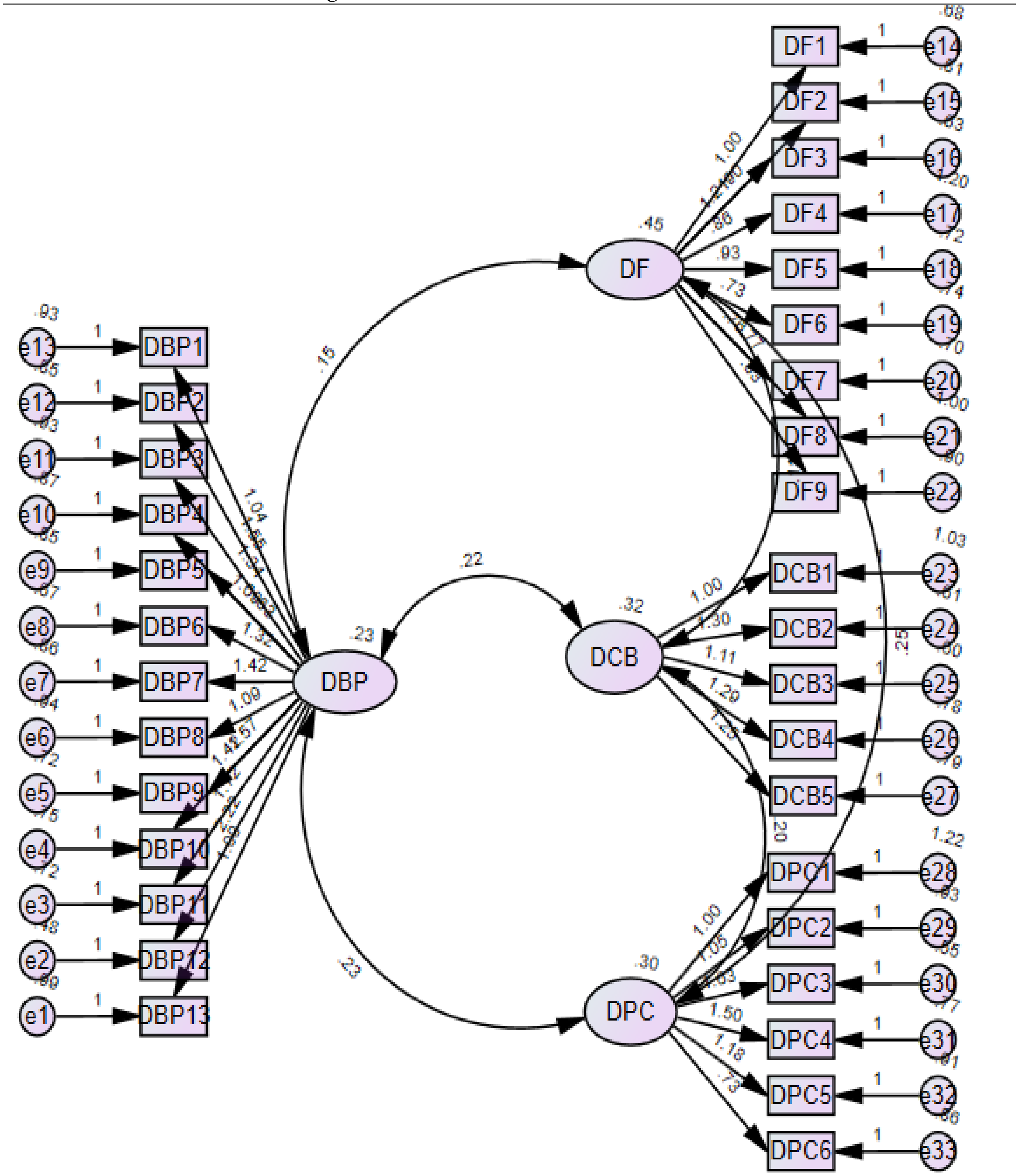


Table 9

Fit Indices of Confirmatory Factor Analysis for FIBS (N=125)

Model	χ^2	*df*	*GFI*	*CFI*	*NNFI*	*RMSEA*
Model Fit	858	489	.69	.75	.73	.08

Note: GFI = Goodness of fit index; CFI = Comparative fit index; NNFI = Normative Fit Index; RMSEA = Root Mean Square error of approximation; χ^2 = chi-square.

Table 9 presents the fit indices for the model fit of the Forman Implicit Bias Scale. The absolute fit of the presented model was χ^2 (489) = 858.885, $p < .001$. The fit index indicated that this model did not provide a good fit. According to Kline (2016), the chi-square test is highly sensitive to sample size and the number of parameters estimated. However, this was the criterion used to evaluate the model fit. The indices of relative fit, GFI, CFI, TLI, and RMSEA, were observed. Hu and Bentler (1999) recommended that χ^2 / df should ideally be between 0 and 3, RMSEA should be 0.08 or less, and the Comparative Fit Index (CFI), Normed Fit Index (NFI), and Goodness of Fit Index (GFI) values of .9 or higher are said to be a good fit. For the present model, the *χ^2/df was 1.75, and the RMSEA was 0.08, both* within a good range. However, the CFI, NFI, and GFI were below the minimum acceptable value of .9. Therefore, the model was considered a poor fit under the standard criteria.

DISCUSSION

The current study focused on creating an indigenous tool to assess implicit bias in trainee healthcare professionals. Following the traditional guidelines of DeVellis's (2017) model, a four-factor, reliable, and valid scale was developed.

The inductive and deductive methods for item generation in Phase I provided in-depth insights into the construct of implicit bias and how cultural beliefs influence healthcare providers' attitudes and behaviors. We saw, from both the healthcare providers' perspective and the patient's point of view, how medical practice is heavily influenced when encountering people from various sociodemographic backgrounds. The scale has been heavily contextualized in Pakistan, where everyday language was used to make it more relatable to test-takers.

Following the focus groups, items that were more redundant than the acceptable limit, irrelevant to their profession, or likely to cause respondent fatigue and incorrect answers were removed to evaluate their overall response to the initial scale. This refinement process was essential to ensure that only items centered on the concept of implicit bias were included in the scale.

Phase II of the research primarily involved Exploratory Factor Analysis. This phase revealed that the scale should have a four-factor structure accounting for approximately 35% of the total variance. The four factors were Dismissive Behavior of Practitioners, Diagnostic Framing, Differential Care Behavior, and Devaluation of Patient Credibility. The first factor demonstrated hidden biases. Kelly's Attribution Theory (1967) supports the idea that practitioners' automatic attributions toward minority patient groups hinder their ability to make compassionate treatment decisions. We also see that practitioner behaviors driven by implicit judgments have previously been shown to affect the quality of care provided (Fitzgerald & Hurst, 2017).

The second factor involves cognitive schemas, in which decisions about diagnosis and treatment are influenced by implicit biases. Many clinicians make these decisions based on deeply rooted cultural perceptions. Maina et al. (2018) found that clinicians' implicit biases are directly linked to disparities in diagnosis and treatment. Kabir and Zaidi (2022) also found that implicit biases lead to misdiagnosis or under-treatment of BIPOC patients. Additionally, Kelly's (1967) theory highlights ingrained perceptions that certain disadvantaged patients, such as those with low socioeconomic status, are seen as non-adherent to medical advice. The third factor assesses visible discriminatory behaviors by practitioners toward patients who are disadvantaged on the basis of sociocultural background, age, or gender. These behaviors appear as differences in how healthcare providers communicate, make eye contact, or show empathy. The last factor concerns skepticism toward patients' narratives, especially among minority groups. This bias is explained by Kelly's (1967) theory, which posits exaggeration or dishonesty when linguistic differences exist between the practitioner and the patient.

Phase III was designed to establish the psychometric properties of FIBS. The scale demonstrated substantial concurrent validity, with convergent and discriminant evidence. The Differential Care Behavior subscale and the BAS simplification subscale showed a moderate positive correlation, consistent with previous research indicating that implicit bias and cognitive simplifications are correlated in clinical settings (Fitzgerald & Hurst, 2017). However, no significant results were found for discriminant validity when comparing the Conscientiousness subscale of the BFI with the FIBS. A possible reason could be the small sample size used for this purpose. Additionally, the self-reports of the two scales may have been affected by social desirability. Regarding criterion validity, nurse trainees showed much higher implicit bias than postgraduate residents, suggesting that implicit bias may be influenced by occupational category. FIBS showed good reliability, as evidenced by high internal consistency and strong temporal stability.

To establish cut-off scores using ROC curve analysis, groups were formed based on low and high implicit bias, defined by percentiles. There was no known diagnostic group as is commonly used. There was no composite cut-off score as circularity was a concern. Although this is a limitation of the present study, the subscale cut-off scores provided an efficient means of classifying FIBS as a diagnostic tool for assessing implicit bias in healthcare settings. Conclusively, Phase III successfully validated the psychometric properties of FIBS, as evidenced by good construct validity, strong reliability, good criterion-related validity, and useful cut-off scores.

CFA was used in Phase IV to empirically validate the newly developed FIBS. However, the model's fit was limited as important fit indices such as Comparative Fit Index (CFI), NFI, and GFI were way below the advised cutoff of .90. This can be attributed to the very small sample size, as a minimum of 200 is highly recommended for worthy CFA results (Kline, 2016). There were practical limitations as data collection had to be terminated early due to a lack of time and resources. To properly confirm the internal structure of the FIBS, future research with a powerful sample is strongly encouraged.

CONCLUSION

Despite certain limitations that can be addressed through further research, there is a significant advance in measuring implicit biases in the healthcare system. For too long, patient care has been kept at the back end in lieu of fulfilling departmental and societal constraints. Measuring, identifying, and mitigating implicit biases among healthcare professionals can lead to much more equitable, compassionate, and unbiased healthcare provision for people from all demographics.

LIMITATION AND STUDY FORWARD

The CFA was done with an extremely low sample size, even failing to meet the minimum 200 requirement, which diminishes the scale's structural validity. The main study included trainee healthcare professionals, excluding those who had been in the field for an extended period. There was moderate convergent and low discriminant validity that required additional assessment to further the scale's construct validity. Lastly, relying solely on the scale is not recommended, as self-report bias is high in such situations due to social desirability, which can compromise the scale's integrity. The FIBS has provided a culturally grounded diagnostic tool to assess implicit bias among trainee healthcare professionals in Pakistan. It can be used in healthcare institutions at training levels to mitigate cultural perceptions related to certain disadvantaged groups, thereby improving patient-provider interactions and reducing healthcare disparities. The subscale-specific cut-off scores can help screen individuals who warrant more scrutiny and reduce their preconceived notions before they enter the practical field. The four-factor structure provides a complex framework that highlights the underlying mechanisms of implicit bias and can inform policy development in healthcare training.

Acknowledgments: The authors appreciate the reviewers' helpful remarks and suggestions for improving our manuscript.

IEER Institute for Excellence in Education & Research

Statements and Declarations
Conflict of interest: We have no conflicts of interest to disclose.
Funding: This work received no funding.
Ethics Statement: The ethical approval was granted by the Forman Christian College (A Chartered University)'s Ethical Review Committee. The reference number is PSYC 699-24-0103.

Informed Consent: Informed consent was obtained from the participants' educational and health institutions. Once approved, informed consent was obtained from participants using a consent form attached to the beginning of the research questionnaire.

Data Availability: The data is available on the Open Science Framework (hyperlink attached).

Author Contributions: Aiza Humayun was responsible for conceptualization, data curation, formal analysis, investigation, methodology, project administration, resources, software, validation, visualization, writing (original draft), and writing (reviewing & editing). Prof. Dr. Sarah Shahed supervised the research.

Word Count: 5010 words (Inclusive of tables/figures).

References

Abbott, S., Shuttlewood, E., Flint, S. W., Chesworth, P., & Parretti, H. M. (2023). "Is it time to throw out the weighing scales?" Implicit weight bias among healthcare professionals working in bariatric surgery services and their attitude towards non-weight focused approaches. *eClinicalMedicine*, *55*, 101770. https://doi.org/10.1016/j.eclinm.2022.101770

Barnes-Holmes, D., Barnes-Holmes, Y., Power, P., Hayden, E., Milne, R., & Stewart, I. (2006). Do you really know what you believe? Developing the Implicit Relational Assessment Procedure (IRAP) as a direct measure of implicit beliefs. *The Irish Psychologist, 32*(7), 169–177.

Bartlett, M. S. (1951). The effect of standardization on a chi-square approximation in factor analysis. *Biometrika,* *38*(3/4), 337–344. https://doi.org/10.1093/biomet/38.3-4.337

Blair, I. V., Steiner, J. F., & Havranek, E. P. (2011). Unconscious (implicit) bias and health disparities: Where do we go from here? The Permanente Journal, 15(2), 71–78. https://doi.org/10.7812/TPP/10-051

Bruce, M., & Weinraub, D. (2023). Implicit gender bias in the clinical judgment of psychopathy and personality disorders among licensed psychologists in the USA. *Journal of Personality Assessment*, *105*(6), 763–769. https://doi.org/10.1080/00223891.2023.2178928

Chapman, E. N., Kaatz, A., & Carnes, M. (2013). Physicians and implicit bias: How doctors may unwittingly perpetuate health care disparities. *Journal of General Internal Medicine,* *28*(11), 1504–1510. https://doi.org/10.1007/s11606-013-2441-1

Chipidza, F. E., Wallwork, R. S., & Stern, T. A. (2015). Impact of the doctor-patient relationship. *The Primary Care Companion for CNS Disorders,* *17*(5), 27354. https://doi.org/10.4088/PCC.15f01840

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research, and Evaluation, 10*(1), Article 7. https://doi.org/10.7275/jyj1-4868

IEER Institute for Excellence in Education & Research

Cvencek, D., Greenwald, A. G., Brown, A. S., Gray, N. S., & Snowden, R. J. (2010). Faking of the Implicit Association Test is statistically detectable and partly correctable. *Basic and Applied Social Psychology*, *32*(4), 302–314. https://doi.org/10.1080/01973533.2010.519236

DeVellis, R. F. (2017). *Scale development: Theory and applications* (4th ed.). SAGE Publications.

FitzGerald, C., & Hurst, S. (2017). Implicit bias in healthcare professionals: A systematic review. *BMC Medical Ethics, 18*(1), 19. https://doi.org/10.1186/s12910-017-0179-8

Fowler, F. J. (2009). *Survey research methods* (4th ed.). Sage. https://doi.org/10.4135/9781452230184

Gonzalez, C. M., Grochowalski, J. H., Garba, R. J., Bonner, S., & Marantz, P. R. (2021). Validity evidence for a novel instrument assessing medical student attitudes toward instruction in implicit bias recognition and management. *BMC Medical Education, 21*, Article 205. https://doi.org/10.1186/s12909-021-02640-9

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review, 102*(1), 4–27. https://doi.org/10.1037/0033-295X.102.1.4

Greenwald, A. G., McGhee, D. E., & Schwartz, J. K. L. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology, 74*(6), 1464–1480. https://doi.org/10.1037/0022-3514.74.6.1464

Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods,* 7(2), 191–205. https://doi.org/10.1177/1094428104263675

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling,* 6(1), 1–55. https://doi.org/10.1080/10705519909540118

Kabir, R., & Zaidi, S. T. (2022). *Implicit bias against BIPOC patients in clinical settings: A qualitative review* (Spectra Undergraduate Research Journal, 2(1), 28–46). https://doi.org/10.9741/2766-7227.1014

Kaiser, H. F. (1974). An index of factorial simplicity for use in factor analysis. *Psychometrika,* 39(1), 31–36. https://doi.org/10.1007/BF02289447

Kelley, H. H. (1967). Attribution theory in social psychology. *Nebraska Symposium on Motivation, 15,* 192–238.

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). The Guilford Press.

Lawrence, B. J., Kerr, D., Pollard, C. M., Theophilus, M., Alexander, E., Haywood, D., & O'Connor, M. (2021). Weight bias among health care professionals: A systematic review and meta-analysis. *Obesity*, *29*(11), 1802–1812. https://doi.org/10.1002/oby.23266

Lovibond, S. H., & Lovibond, P. F. (1995). *Manual for the Depression Anxiety Stress Scales* (2nd ed.). Psychology Foundation of Australia.

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research,* *35*(6), 382–385. https://doi.org/10.1097/00006199-198611000-00017

IEER Institute for Excellence in Education & Research

Maina, I. W., Belton, T. D., Ginzberg, S., Singh, A., & Johnson, T. J. (2018). A decade of studying implicit racial/ethnic bias in healthcare providers using the implicit association test. *Social Science & Medicine, 199*, 219–229. https://doi.org/10.1016/j.socscimed.2017.05.009

McConahay, J. B. (1986). Modern racism, ambivalence, and the modern racism scale. In J. F. Dovidio & S. L. Gaertner (Eds.), *Prejudice, discrimination, and racism* (pp. 91–125). London: Academic Press.

O'Shea, B., Watson, D. G., & Brown, G. D. A. (2016). Measuring implicit attitudes: A positive framing bias flaw in the Implicit Relational Assessment Procedure (IRAP). *Psychological Assessment, 28*(2), 158–170. https://doi.org/10.1037/pas0000172

Ridd, M. J., Shaw, A., Lewis, G., & Salisbury, C. (2009). The patient–doctor relationship: A synthesis of the qualitative literature on patients' perspectives. *British Journal of General Practice, 59*(561), e116–e133. https://doi.org/10.3399/bjgp09X420248

van Ryn, M., & Fu, S. S. (2003). Paved with good intentions: Do public health and human service providers contribute to racial/ethnic disparities in health? *American Journal of Public Health, 93*(2), 248–255. https://doi.org/10.2105/AJPH.93.2.248

Watts, L. L., Medeiros, K. E., McIntosh, T. J., & Mulhearn, T. J. (2020). *Decision biases in the context of ethics: Initial scale development and validation. Personality and Individual Differences, 153*, Article 109609. https://doi.org/10.1016/j.paid.2019.109609

Yasmeen, S., Nazar, K., & Perveen, T. (2023). Implicit bias in health care providers during providing health care services to patients in a tertiary care hospital South Punjab Pakistan. *Basic and Applied Nursing Research Journal*, *4*(1), 40–45

Yzerbyt, V. Y., Rogier, A., & Fiske, S. T. (1998). Group entitativity and social attribution: On translating situational constraints into stereotypes. *Personality and Social Psychology Bulletin, 24*(10), 1089–1103. https://www.researchgate.net/publication/288802213_Development_and_psychometric_evaluation_of_ageism_attitude_scale_among_the_university_students

IEER Institute for Excellence in Education & Research