Charlotte Fresenius Hochschule

Studiengang: Psychologie (B. Sc.)

Studienort: München

Bachelorarbeit im Studiengang B.Sc.

„Clinimetric Properties of the German Version of the Euthymia Scale (ES): Validity and Sensitivity Analysis “

vorgelegt von:

Nico Andre Steffen

(Matr. -Nr.: 400334811)

6. Fachsemester

Erstgutachter: Prof. Dr. Stephan Goerigk

Zweitgutachterin: Dr. Fabienne Große-Wentrup

Abgabedatum: 15.07.2025

1) Table of Contents

2) List of Figures

3) List of Tables

4) List of Abbreviations

Abstract

Background

Deutsch und Englisch!

Prereg.

Introduction

Over the past two decades the importance of well-being has been increasingly acknowledged (Blanchflower & Oswald, 2011; Giovanni A. Fava & Bech, 2016; Hicks et al., 2013; Naci & Ioannidis, 2015). Well-being is a key component of the World Health Organizations' definition of mental health and therefore a crucial aspect of health in general (Organization & Others, 2021). While there is much agreement on the general importance of well-being, there are fundamental differences in definition (Dodge et al., 2012) and theoretical basis (Deci & Ryan, 2008). Across disciplines (i.e., public health, clinical needs, politics, health economics) there are different priorities as to what well-being should measure (Diener et al., 2010). In the research of well-being there are two main perspectives: the hedonistic tradition defines well-being as feeling happy or showing high positive affect and low negative affect. It focusses on maximizing pleasure and minimizing pain. The term subjective well-being (SWB) (Diener, 1984), a widely used operationalization of well-being, originates from the hedonic tradition. Eudaimonia on the other hand has da deeper and more complex understanding of well-being. Dating back to Aristotle's "Nicomachean Ethics" (Irwin, 2019) the eudaimonic tradition views well-being as fulfilling one's true potential, fulfilling meaningful goals and self-actualization (Deci & Ryan, 2008). Psychological well-being with measurement scales like the psychological well-being scale (PWB) (C. D. Ryff & Keyes, 1995; Carol D. Ryff, 1989) is rooted in this tradition.

While traditional well-being measures focus on hedonic or eudaimonic perspectives, they often fail to meet clinical needs which differ from those in positive, general, social or developmental psychology. They often present a fragmented and reductionist view of well-being that doesn't reflect the complex nature of well-being.

These frameworks are often disconnected from clinical realities, lacking relevance for individuals with mental health challenges (A. M. Wood & Tarrier, 2010). The clinical consideration of psychological well-being thus required a novel framework (Guidi & Fava, 2022).

Euthymia

Taking on these challenges Fava and Bech (2016) provided a novel definition of euthymia which was discussed in detail in subsequent publications (Giovanni A. Fava & Guidi, 2020a; Guidi & Fava, 2022). With their definition of euthymia they presented a more integrated and comprehensive multidimensional construct of well-being that aligns with the complexities of mental health and better supports clinical interventions.

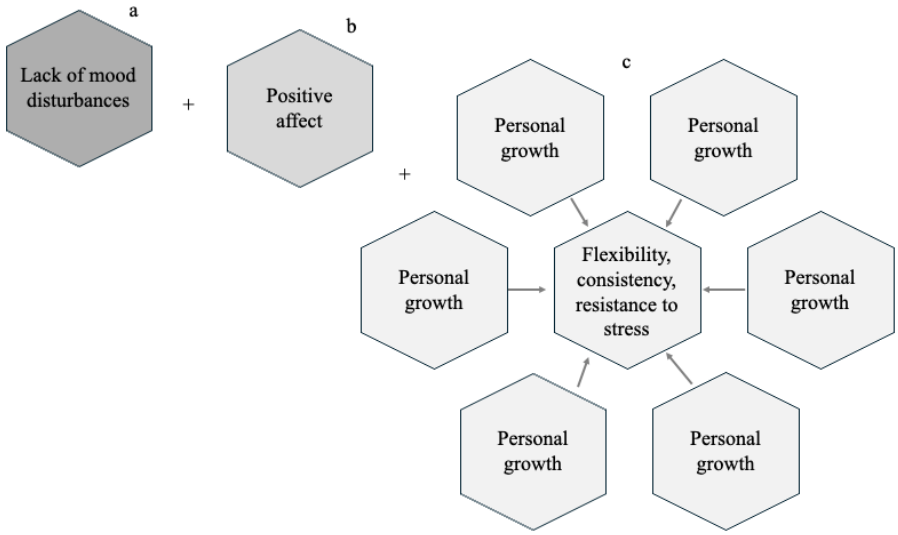
They characterize euthymia by following features (Guidi & Fava, 2022) (Figure 1):

a) A lack of mood disturbances (i.e., diagnostic rubrics): One should be in full remission (if prior mood disorder existed) not experiencing symptoms of clinical significance. Negative affect like sadness or anxiety may still be experienced but should be short lived and not negatively impact everyday life.
b) The presence of positive affect (i.e., feeling cheerful, calm, active, interested in things and experiencing restorative sleep). This dimension overlaps with the concept of subjective well-being (Diener, 1984).
c) The third component encompasses balanced levels of well-being dimensions and integration derived from work by Marie Jahoda (1959): Jahoda identified six dimensions of positive mental health – (1) autonomy, (2) environmental

mastery, (3) positive interactions with others, (4) personal growth, (5) development or self-actualization, and (6) attitude towards oneself. Ryff (1989) later translated these dimensions into a self-rated questionnaire (The Psychological Well-Being scales; PWB) slightly rewording the dimensions. Further, integration was defined by Jahoda as (1) a balance of psychic forces (flexibility), (2) a unifying outlook on life (consistency) and (3) resistance to stress (resilience).

Figure 1

The Unifying Concept of Euthymia as defined by Guidi and Fava (2022)



Existing measures of euthymia include the Euthymia Scale (ES) (Giovanni A. Fava & Bech, 2016) - a 10-item self-report questionnaire, and the Clinical Interview for Euthymia (CIE) (Giovanni A. Fava & Guidi, 2020a) – a 22 item structured interview. These Instruments were developed using clinimetric principles (G. A. Fava et al., 2012; Alvan R. Feinstein, 1987) which will be explained in detail in the next section. Apart from the form of administration (questionnaire vs. structured interview) the two instruments differ in the amount of items: The Euthymia Scale (ES) consists of five questions adopted from the WHO-5 well-being index (Topp et al., 2015) reflecting

point b (presence of positive affect) of the displayed euthymia model (Figure 1) and five questions addressing the individual's balance among psychic forces leading to high levels of resilience and frustration tolerance (point c). The Clinical Interview for Euthymia (CIE) expands on these 10 questions, adding 12 questions derived from the Psychological Well Being Scale (PWB) (Carol D. Ryff, 1989) – each well-being dimension being represented by two questions – providing a more nuanced perspective on point c.

Up to this date, the Euthymia Scale (ES) has not been validated within a German speaking population. Therefore, it is crucial to perform a clinimetric analysis for the German version of the Euthymia Scale (ES-G).

Clinimetrics

The term clinimetrics was first introduced by Feinstein (1987) referring to the development and use of rating scales, indexes, and instruments measuring clinical phenomena that cannot be measured using traditional laboratory methods. As an early example for clinimetric measures he mentioned the Apgar Score (Apgar, 1953) evaluating a newborn infants' health condition. Feinstein shed light on the lack of standards for rating scales within clinical use and highlighted the conflict between the scientific goal of standardization (reliability and validity) and the clinical goal of sensibility (face validity, content validity and ease of use). Criteria for the development of clinimetric rating scales were described (A. R. Feinstein, 1983; Alvan R. Feinstein, 1987; Jones & Feinstein, 1982) and further refined in a subsequent publication (Wright & Feinstein, 1992).

The clinimetric approach, also referred to as the science of clinical measurements (G. A. Fava et al., 2012) therefore provides a set of guidelines for the development and validation of existing patient-reported outcome measures (PROMs) aligning with clinical goals and patients' needs, which the more common psychometric approach often misses to address (Wright & Feinstein, 1992).

There are several differences between the clinimetric and psychometric approaches: historically the development of psychometrics took place in research fields outside of clinical psychology, mainly in educational or social sciences (Giovanni A. Fava et al., 2004; Wright & Feinstein, 1992) while clinimetrics was developed specifically for measuring clinical phenomena (Alvan R. Feinstein, 1987). Regarding the selection of items the focus of the psychometric framework is often laid on homogeneity- referring to a high degree of inter-item correlations – leading to a set of items that essentially all measure the same thing (Bech, 2004; G. A. Fava et al., 2012; Tomba & Bech, 2012; Wright & Feinstein, 1992). However, the goal of a high score for homogeneity of components may contradict with clinimetric properties, in particular sensitivity to change (Giovanni A. Fava & Belaise, 2005). This may also lead to the inclusion of redundant items, reducing clinical applicability (Carrozzino, 2019). Thus, following the clinimetric approach, homogeneity and unidimensionality are not of primary interest and items should instead be providing non-redundant, clinically distinct information (Wright & Feinstein, 1992). While psychometrics focusses on construct, convergent, divergent, and criterion validity, clinimetrics emphasizes clinical, predictive, incremental, and biological validity (Carrozzino, Patierno, et al., 2021).

Initiatives like PROMIS (Patient-Reported Outcomes Measurement Information System) (D. Cella et al., 2007; David Cella et al., 2010; Rothrock et al., 2011)

or COSMIN (Consensus-based Standards for the selection of health Measurement Instruments) (L. B. Mokkink et al., 2018, 2006; Lidwine B. Mokkink et al., 2016, 2010) often build the foundational framework in the development and validation of PROMs and are strongly rooted in the psychometric tradition. It is questionable if these frameworks are suited for complex clinical realities.

Carrozzino et al. (2021) present a comprehensive overview of the methodological differences between psychometrics and clinimetrics in the context of reliability and validity testing of PROMs and provide recommendations for the analysis of clinimetric patient-reported outcome measures (CLIPROM criteria). Important CLIPROM criteria are:

Sensitivity

The concept of sensitivity refers to the ability of a rating scale (or single items of a rating scale) or self-report questionnaire to differentiate between different groups of subjects (e.g., patients and healthy controls, depressed inpatients or outpatients) and to reflect outcome changes in clinical trials (Kellner, 1972). In this context, a clinimetric rating scale should also be able to differentiate between groups receiving therapeutic intervention and placebo or attention control groups (Giovanni A. Fava et al., 2018). If clinical trials fail to differentiate between these groups the reason may be poor performance of the treatment, but in some cases it might be due to a lack of sensitivity of the used outcome measures (Giovanni A. Fava et al., 2004). The sensitivity of a rating scale is a crucial criterion for their use in clinical routines.

Validity

Clinical validity. Refers to the ability of a measure to accurately identify or discriminate subjects with or without a specific condition (i.e., depression vs. no depression) (Carrozzino, 2019; Carrozzino, Christensen, & Cosci, 2021; Giovanni A. Fava et al., 2004; A. Feinstein, 1987). In comparison to the criteria of sensitivity, which is about detecting meaningful differences in treatment effects, clinical validity is specifically about accurate diagnostic discrimination (i.e., correctly identifying presence or absence of a condition).

Construct validity. The concept of construct validity was first introduced by Cronbach and Meehl (1955), and refers to how well a rating scale measures the underlying theoretical concept it is intended to measure (Strauss & Smith, 2009). Following psychometric guidelines, it is often assessed via factor or principal component analysis. But the utility of these methods for the clinical use has been questioned (Bech, 2012; Giovanni A. Fava et al., 2018; Alvan R. Feinstein, 1987): Psychometric models reveal structure, but do not guarantee that the total score reflects the severity of a clinical condition (Bech, 2012). In the clinimetric approach, unidimensionality of an instrument is not of primary interest (Wright & Feinstein, 1992). In clinimetric analyses, construct validity can be assed though methods like Rasch and Mokken analyses (Bech, 2012; Carrozzino, Christensen, & Cosci, 2021; Mokken, 1970; Rasch, 1993), evaluating the extent to which items provide distinctive clinical information and symptoms represented by a clinimetric scale belong to an underlying clinical syndrome (Bech, 2012; Carrozzino, Christensen, & Cosci, 2021).

Predictive validity. Refers to the ability of a rating scale to predict future outcomes like treatment response (i.e., responder vs. non-responder) or psychological distress scores after a certain period of time (Carrozzino, Patierno, et al., 2021).

Incremental validity. Indicating that a rating scale - or each item of a scale - should add meaningful information beyond what is already available through other accessible information (Sechrest, 1963). Incremental validity can be assessed through hierarchical regression analyses.

Concurrent validity. Concurrent validity refers to the degree to which a measurement tool correlates with existing, previously validated instruments (Bagby et al., 1994). But a high correlation between two instruments alone does not indicate good validity of the instrument: The scales may measure a common aspect but still differ in clinical validity or sensitivity. Thus, concurrent validity in clinimetric analyses is not considered as important as other criteria (Giovanni A. Fava et al., 2004).

The present study

The aim of this study was to create a German translation of the Euthymia Scale (ES) and to validate the ES-G through a comprehensive clinimetric analysis. The analysis plan was designed in adherence to the recommendations for clinimetric patient-reported outcome measures (CLIPROM) as outlined by Carrozzino et al. (2021). Additionally, the performance of a self-created 6-point Likert version of the ES-G was tested against the original dichotomous version.

Research Objectives

The following research objectives were addressed. After each objective it is stated in brackets, to which of the above mentioned CLIPROM criteria it corresponds to.

1. Rationale for the German translation of the ES
2. Correlation Analysis (*concurrent validity*)

3. Rasch analysis (*construct validity / dimensionality*)
4. Ability of the ES-G to predict whether a patient will be a responder or non-responder to psychotherapy (*predictive validity*)
5. Ability of the ES-G to predict whether a subject is clinical or non-clinical (*sensitivity*)
6. Ability of the ES-G to reflect symptom changes in psychotherapy (*sensitivity*)
7. Ability of the ES-G to discriminate between healthy subjects and subjects with a past or current depression (*clinical validity*)
8. Ability of the ES-G to discriminate between symptom severity groups (*clinical validity*)
9. Determining a cutoff score for differentiating subjects with or without depression
10. Incremental validity of the ES-G (*incremental validity*)
11. Comparison of the self-adapted 6-point Likert version of the ES-G with the original version

Hypotheses for concurrent validity

The following a priori hypothesis for concurrent validity were postulated: The correlation between the ES-G and …

H1: psychological distress was expected to be negative

H2: quality of life was expected to be positive

H3: trait resilience was expected to be positive

H4: psychological well-being was expected to be positive

H5: depressive symptoms was expected to be negative

Methods

Study Design

This study utilized data from two sources: (1) a clinical feasibility trial, evaluating the transdiagnostic Well-Being Therapy (WBT) (G. A. Fava, 2016) in a group therapy format at the day clinic of the LMU Hospital in Munich, and (2) a cross-sectional online survey targeting non-clinical participants. This design allowed for cross-sectional and longitudinal analyses.

Participants

Non-clinical participants were eligible for inclusion if they (1) were between 18 and 75 years old, (2) spoke German fluently, and (3) provided informed consent. Exclusion criteria were: (1) the presence of inadequately treated concomitant somatic disease (e.g., current hypothyroidism and hypertension) including acute and chronic infections or autoimmune diseases, and (2) pregnancy or breastfeeding. After clearing for exclusion criteria (n = 16 somatic disease, n = 2 pregnant, and n = 2 age over 75) a total of $N = 181$ non-clinical participants (146 females [81%]; $M = 25.36$ years, $SD = 8.88$) were included in the study.

Day clinic patients were eligible if they met the same general criteria (age, language, consent, pregnancy, untreated somatic disease), and additionally: (1) were not acutely suicidal, and (2) did not have a primary diagnosis of organic mental disorder (F00-F09), mental and behavioral disorder due to psychoactive substance use (F10-F19; F63), or eating disorder (F50). Eligible patients were diagnosed with at least one of the following psychiatric conditions: affective disorder (F30 – F39), schizophrenia, schizotypal, delusional, and other psychotic disorder (F20 – F29), anxiety disorder

(F40 – F41), obsessive-compulsive disorder (F42), dissociative, stress-related, somatoform and other nonpsychotic mental disorder (F43 – F48), or personality disorder (F60 – F62).

As of the current data cutoff, 32 patients (19 females [59%]; *M* = 39.09 years, *SD* = 13.25) had completed baseline assessment (t0) and a total of 25 patients completed both pre- (t0) and post (t1)-assessment.

Based on literature recommendations (Charter, 1999; Frost et al., 2007; Schönbrodt & Perugini, 2013) a minimum sample size of *N* = 238 (119 patients and 119 non-clinical) was pre-registered to ensure stable estimates of reliability and validity. Detailed descriptions of the final sample characteristics are displayed in Tables 2 and 3.

Procedure

The two samples were recruited through separate procedures and assessed using different formats. Recruitment methods and data collection procedures are outlined below.

Non-clinical participants were recruited between October 2024 and May 2025 through study flyers and presentations in university lectures. They were invited to participate in a cross-sectional online survey through the platform Unipark (Tivian XI GmbH).

Day clinic patients were recruitment at the LMU Hospital between August 2024 and May 2025 as part of an ongoing feasibility trial. This sample underwent an eight-week multimodal therapy program that included Well-being Therapy (WBT) in both group and individual formats. Data from this sample were collected at two time

points: t0 (upon admission) and t1 (at discharge after eight weeks). In the clinical sample, questionnaires were administered in a paper-and-pencil format.

This study was preregistered on the Open Science Framework (OSF; https://osf.io/yr8e5/?view_only=c9ddd629046148068bfbfdaab219e27a) and received approval from the Ethics Committee of the LMU (Faculty of Medicine, LMU Munich, Munich, Germany, project-no.: 24-0359). All deviations from the preregistration are transparently reported in the discussion. Informed consent was obtained from all participants prior to their inclusion in the study.

Measures

Euthymia Scale (ES)

The Euthymia Scale (ES) (Giovanni A. Fava & Bech, 2016) is a 10 – item self-report clinimetric measure. All Items are scored dichotomously as 1 (true) or 0 (false). Items 6 – 10, measuring psychological well-being, were adopted from the World Health Organization-5 Well-Being Index (WHO-5) (Topp et al., 2015). Items 1 – 5 measure levels of psychological flexibility. While Fava & Bech (2016) recommend calculating a global euthymia score, ranging from 0 – 10, with higher scores indicating higher levels of euthymia, Carrozzino et al. (2019) suggest a two dimensional structure and recommend using separate scores for the two subscales. Clinimetric analyses of the Japanese (Sasaki et al., 2021; Sasaki & Nishi, 2022) and Italian (Carrozzino et al., 2019) versions have shown, that the Euthymia Scale (ES) is a valid and highly sensitive clinimetric index. For the present study an adapted 6-point Likert version (from 0 “at no time” to 5 “all of the time”) was used in addition to the original format. The scale format was adapted from the WHO-5. Both scales were administered in a German version (ES-G), and total sum scores were calculated.

Beck Depression Inventory II (BDI-II)

The Beck Depression Inventory II (BDI-II) (Beck et al., 1996) is a widely used self-report instrument for assessing the severity of depressive symptoms in clinical and non-clinical populations. It is based on the diagnostic criteria for major depressive disorder as outlined in the DSM-IV (American Psychiatric Association et al., 1994). The BDI-II consists of 21 items, each representing a symptom related to depression. Items are rated on a 4-point Likert scale ranging from 0 (no symptom) to 3 (severe symptom), resulting in a total score between 0 and 63. For the German version (Hautzinger et al., 2006), internal consistency (Cronbach's α) was reported as good ($\alpha \geq .84$) (Kühner et al., 2007). The BDI-II differentiates well between different severity levels of depression and is sensitive to change. In this study, cutoff scores were interpreted as recommended by Beck et al. (1996): minimal depression (0-13), mild depression (14-19), moderate depression (20-28), and severe depression (29-63).

World Health Organization Quality of Life (WHOQOL-BREF)

The WHOQOL-BREF (Group & Others, 1998) is a self-report questionnaire developed by the World Health Organization (WHO) to assess individuals' subjective quality of life. It is derived from the original WHOQOL-100 and consists of 26 items. It measures four domains: (1) Physical health, (2) psychological, (3) social relationships, and (4) environment. In addition to the domain scores, two items assess overall quality of life and general health. Items are rated on a 5-point Likert scale, with higher scores indicating better quality of life. Items 3, 4, and 26 are negatively worded and need to be reverse-scored. The internal consistency (Cronbach's α) of the four domains was reported between .57 and .88 for the German version of the WHOQOL-BREF (Angermeyer et al., 2000). In this study domain scores were converted to a 0-100 scale,

as recommended by the authors. A mean total score was then calculated by averaging the four domain scores, providing an overall index of subjective quality of life.

Psychological Well-Being Scale (PWB-18)

The 18-item version of the Psychological Well-Being Scale (PWB) (C. D. Ryff & Keyes, 1995) is a short form of the original 84-item instrument developed by Ryff (1989). The PWB measures six theoretically grounded dimensions of psychological well-being based on Jahoda (1959): (1) autonomy, (2) environmental mastery, (3) personal growth, (4) positive relations with others, (5) purpose in life, and (6) self-acceptance. Each dimension is assessed by three questions, rated on a 6-point Likert scale ranging from 1 (strongly disagree) to 6 (strongly agree). Eight items need to be reverse-coded. Previous studies have reported low internal consistencies for the 18-item version, with Cronbach's α ranging from .33 to .56 (C. D. Ryff & Keyes, 1995). In the present study, both dimensional scores (range: 3 – 18) and a total psychological well-being score (range: 18 – 108) were used.

Connor Davidson Resilience Scale (CD-RISC-10)

The Connor-Davidson Resilience Scale (CD-RISC) is a widely used self-report measure for assessing trait resilience, defined as the ability to cope well with stress and adversity. The original scale consists of 25 items (Connor & Davidson, 2003), but a 10-item short version (Campbell-Sills & Stein, 2007) has been validated and is commonly used. The CD-RISC-10 includes 10 items rated on a 5-point Likert scale ranging from 0 (not true at all) to 4 (true nearly all the time), with total scores ranging from 0 to 40. The German version has shown good internal consistency (Cronbach's α = .84) and test-retest reliability (r_{tt} = .81) (Sarubin et al., 2015).

Brief Symptom Inventory (BSI-53)

The Brief Symptom Inventory (BSI-53) (Derogatis, 1993; Franke & Derogatis, 2000) is a self-report measure to assess psychological symptom burden across a wide range of psychiatric dimensions: (1) Somatization, (2) obsessive-compulsive, (3) interpersonal sensitivity, (4) depression, (5) anxiety, (6) hostility, (7) phobic anxiety, (8) paranoid ideation, and (9) psychoticism. The BSI-53 contains of 53 Items, each rated on a 5-point Likert scale ranging from 0 (not at all) to 4 (extremely), reflecting symptom distress over the past 7 days. In addition to the domain scores, three global indices can be calculated: The Global Severity Index (GSI), the Positive Symptom Distress Index (PSDI), and the Positive Symptom Total (PST). The BSI-53 has shown good psychometric properties, including high internal consistency with Cronbach's α for the GSI typically exceeding .90 (Endermann, 2005). In the present study, the Global Severity Index (GSI), calculated as the mean score of all items, was used as a general measure for psychological distress.

WHO-5 Well-Being Index

The WHO-5 Well-Being Index (Health Organization, 1998) is one of the most commonly used self-report rating scales for assessing subjective well-being (SWB) in research and clinical settings. The five questions are rated on a 5-point Likert scale ranging from 0 (at no time) to 5 (all of the time), resulting in a raw score range of 0 to 25. For better comparison with other well-being measure, the raw score is typically multiplied by four, resulting in a percentage score from 0 to 100. The WHO-5 has demonstrated high clinimetric validity, can be used as an outcome measure, and serves as a screening tool for depression. It has shown high internal consistency across various studies with Cronbach's α typically exceeding .80 (Topp et al., 2015).

Mini-International Neuropsychiatric Interview for Depression (MINI)

The Mini-International Neuropsychiatric Interview (MINI) is a brief, structured diagnostic interview developed to assess the presence of DSM-IV or ICD-10 psychiatric disorders (Sheehan et al., 1998). In the present study, only the Major Depressive Episode (MDE) module was used, adapted as a self-reported format, to assess the presence of current and past depressive episodes. The MDE module consists of nine dichotomous items (yes/no), each representing a symptom based on DSM-IV criteria for diagnosing depression. Participants were classified as having a current MDE, past MDE, or nor lifetime MDE. For past MDE, participants were categorized as YES (endorsed 5-9 symptoms) or NO (0-4 symptoms). For current MDE, three categories were used: MDE (5-9 symptoms), subthreshold depression (1-4 symptoms), or none (0 symptoms). This grouping approach was adopted from Sasaki et al. (2021).

Translation of the Euthymia Scale

To create a German version of the Euthymia Scale (ES-G), items 6 to 10 were adopted from the official German translation of the WHO-5 (Health Organization, 1998). The remaining five items were derived from an existing version used in a published translation of the Clinical Interview for Euthymia (CIE) (Giovanni A. Fava & Guidi, 2020b).

Unlike the Euthymia Scale (ES), the Clinical Interview for Euthymia (CIE) uses negatively worded items. Therefore, item 2 of the ES (“I do not keep thinking about negative experiences”) required a slight rewording compared to its counterpart in the CIE (“Do you keep thinking of negative experiences”). The final version of the Euthymia Scale (ES-G) is presented in Appendix A, Table A1.

Statistical analyses

All statistical analyses were carried out using R 4.5.0 (R Core Team, 2023). The alpha level for statistical significance was $\alpha = .05$. Descriptive statistics were calculated to summarize participant characteristics and key study variables. For continuous variables, means and standard deviations were reported, for categorical variables, frequencies and percentages were calculated. Between-group comparisons were performed using Mann-Whitney U tests, Fisher's exact tests and *t*-tests, depending on data type and distribution.

Model assumptions for all parametric tests (e.g., normality of residuals, homoscedasticity) were examined prior to conducting analyses and are provided in Appendix C.

Missing data was handled as follows: If missing items exceeded 10% for a questionnaire, the participant was excluded. With ≤ 10% mean imputation was applied.

Concurrent validity

To assess concurrent validity (**Objective 2**) of the ES-G, Spearman rank correlation analyses were conducted between the ES-G total score and related constructs, including psychological distress (GSI), quality of life (WHOQOL-BREF), resilience (CD-RISC), psychological well-being (PWB), and depressive symptoms (BDI-II). P-values were corrected for multiple comparisons using the False Discovery Rate (FDR) procedure (Benjamini & Hochberg, 1995).

Construct validity / Dimensionality

To evaluate the dimensionality (**Objective 3**) of the ES-G, Rasch analysis was performed using the easyRasch package (Johansson, 2025a). This analysis was guided

by recommendations from Johannson et al. (2023) with a focus on the following indicators of dimensionality:

Item Fit. Was assessed using conditional infit statistics, which are robust to sample size and preferred over traditional unweighted mean square (outfit) or z-standardized fit statistics (ZSTD) values (Johansson, 2025b; Müller, 2020). Infit is an information-weighted mean square residual, which reflects the degree to which observed item responses align with expected responses under the Rasch model. Information weighted mean square (InfitMSQ) was calculated by multiplying the squared standardized residuals by the observed response variance and then divided by the sum of the item response variances. Values substantially above or below 1.0 may indicate item misfit. To determine item-specific cutoff values, a parametric bootstrap procedure with 200 iterations was conducted, in line with the recommendations by Johansson (Johansson, 2025b).

Principle Component Analysis of item residuals (PCAR). While earlier rules of thumb suggested a cutoff of 1.5 for the first eigenvalue (Smith, 2002) to support unidimensionality, later research has shown that the expected PCAR eigenvalue also depends on sample size and test length (Chou & Wang, 2010). Therefore, a simulation-based approach was used to estimate a more appropriate cutoff for the first eigenvalue in this sample. As recommended by Johansson (2025a), the distribution of eigenvalues was simulated with a parametric bootstrap procedure, using 500 iterations to determine a cutoff value for the largest PCAR eigenvalue.

Local independence. According to the Rasch model, items should be locally independent, meaning they should only correlate through the latent trait. Violations of

this assumption may indicate redundancy, item clustering, or multidimensionality. Local independence was therefore assessed by examining residual correlations between item pairs (Kim et al., 2011). To get a useful cutoff threshold for residual correlations a bootstrapping procedure with 400 iterations was conducted as recommended by Christensen et al. (2017). Items with residual correlations above the calculated threshold were considered as locally dependent.

To further validate the results of the Rasch analyses, a parallel analysis based on factor analysis with 1000 iterations to generate simulated and resampled datasets was conducted. The quantile criterion was set at .95.

Predictive validity

Predictive validity refers to the ability of a rating scale to predict future (treatment) outcomes. It was tested if baseline ES-G total scores could predict whether a patient would respond to psychotherapy (**Objective 4**). Response was evaluated using two outcomes: (1) a positive well-being criterion (WHO-5), where patients with an increase of ≥ 10 points from t0 to t1 were considered responders (Topp et al., 2015); and (2) a symptom reduction criterion (BDI-II), where a ≥ 50% change was used to define response.

A machine learning-based predictive modeling approach was employed using logistic regression classifiers implemented with the mlr package (Bischl et al., 2016). Model performance was evaluated using nested cross-validation with 5 folds and 10 repetitions. The models were optimized for balanced accuracy (BAC). Due to imbalanced group sizes, random undersampling was applied within the inner CV loop. This strategy improves robustness of predictive models with imbalanced classification tasks (He & Garcia, 2009).

Sensitivity

To evaluate the sensitivity of the ES-G, two analyses were conducted:

(1) It was tested whether baseline ES-G total scores could predict group membership (non-clinical participants vs. day clinic patients; **Objective 5**). A machine learning-based logistic regression model was trained using the same approach described under predictive validity. Model performance was assessed via nested 5-fold cross-validation (10 repetitions), optimized for balanced accuracy. To address class imbalance, random undersampling was applied within the inner CV loop.

(2) To examine the ES-G's sensitivity to symptom change (**Objective 6**), it was tested whether changes from baseline (t0) to post-treatment (t1) in the BDI-II were associated with changes on the ES-G within the clinical sample. A sandwich linear regression model, with ΔBDI-II as the criterion and centered ΔES-G as the predictor, controlled for centered BDI-II baseline scores was employed:

$$\Delta\text{BDI-II} \sim \Delta\text{ES-G}_{\text{centered}} + \text{BDI-II}_{\text{t0,centered}}$$

Proof of sensitivity to change was defined as a significant Wald test of the ΔES-G (centered) slope, with an expected negative *b* coefficient. To quantify the unique explained variance of both ES-G versions, partial R^2 was reported.

Clinical validity

Due to violations of homogeneity of variances (Appendix C, Table C2), one-way Welch's ANOVAs were conducted to assess the clinical validity of the ES-G. It was tested whether ES-G total scores differed across groups based on (1) depression history (**Objective 7**) and (2) symptom severity (**Objective 8**).

(1) Participants from both samples were classified into five groups based on current and past MDE status, assessed by a self-report version of the Mini-International Neuropsychiatric Interview (MINI) (Sheehan et al., 1998). This grouping strategy was adopted from Sasaki et al. (2021) and is presented in Appendix B, Table B1.

(2) Symptom severity groups were created according to established BDI-II cutoff scores (Beck et al., 1996): *minimal* (0–13), *mild* (14–19), *moderate* (20–28), and *severe* (≥29) depressive symptoms. These groups included participants from both clinical and non-clinical samples.

Jonckheere-Terpstra trend tests with 10,000 permutations were performed to assess whether a decreasing trend in ES-G total scores was observed across ordered groups with increasing symptom burden. Games–Howell post-hoc comparisons were used to account for unequal group variances. Omega squared (ω^2) was used as the effect size measure and estimated using a bootstrapping procedure with 1,000 resamples.

Cutoff determination

To determine a clinically meaningful cutoff score for the ES-G for screening subjects with or without depression, receiver operating characteristics (ROC) curve analysis (Metz, 1978; Zweig & Campbell, 1993) were conducted. As a reference criterion BDI-II scores were used. In their meta-analyses von Glischinski et al. (2019) recommend using different cut points to screen for depression in primary care and healthy populations vs. psychiatric settings. For the non-clinical sample, a BDI-II score of ≥ 13 was used to define depression while for the clinical sample, a score of ≥ 19 served as the cut point, as suggested by von Glischinski et al. (2019). ROC curve analyses were performed for both the original version of the Euthymia Scale and the

adapted 6-point Likert version. Analyses were carried out using the R package pROC (Robin et al., 2011). The following indicators were reported: area under the curve (AUC), sensitivity, specificity. The optimal cutoff scores were determined using Youden's J statistic, which maximizes the sum of sensitivity and specificity.

Incremental validity

Hierarchical linear regression analyses were used to assess incremental validity of the ES-G over the WHO-5. The criterion variable was the Psychological Well-Being Sclae (PWB) total score. Predictors were entered in the following order: WHO-5 at step 1, the ES-G at step 2. An increase in the explained variance (ΔR^2) from step 1 to step 2 was interpreted as an indicator for incremental validity. All models were controlled for sex, age, and education as these demographic variables have been shown to be associated with well-being outcomes (Buecker et al., 2023; Carrozzino et al., 2019; Oishi & Tay, 2019; W. Wood et al., 1989).

Comparison of the Self-Adapted 6-Point Likert Version of the ES-G with the Original Version

The performance of the self-adapted 6-point Likert version of the ES-G was compared to the original version. This comparison was based on balanced accuracy (BAC) scores from the predictive modeling objectives (Objectives 4 and 5), explained variance (R^2) from the sensitivity to change analysis (Objective 6) and hierarchical regression models (Objective 10), and effect sizes (ω^2) from the ANOVA analyses (Objectives 7 and 8).

Maybe: tabelle welche packete in r?

Results

Participants

Descriptive statistics of the final sample (N = 213) are presented separately for sociodemographic characteristics (Table 1) and study variables (Table 2).

The full sample at baseline consisted of 165 female (77.5%), 46 male (21.6%), and 2 participants who identified as divers (0.9%). The mean age of participants was 27.43 years (SD = 10.81).

Statistical analyses revealed significant differences in distribution of categorial variables and mean scores of continuous variables between the clinical and non-clinical sample. Levene's test indicated homogeneity of variances for all comparisons (all $p > .05$). Shapiro-Wilk tests revealed significant deviations from normality in all variables within the non-clinical sample. In the clinical sample only the WHO-5 deviated from normality. However, Mann-Whitney U tests yielded the same pattern of results as the parametric *t*-tests; therefore, only the results of the *t*-tests are reported.

In the clinical sample, primary diagnoses were as follows: major depressive disorder ($n = 21$; 65.6%), borderline personality disorder ($n = 3$; 9.4%), anxiety disorder ($n = 2$; 6.3%), obsessive-compulsive disorder ($n = 2$; 6.3%), autism spectrum disorder ($n = 1$; 3.1%), and schizophrenia ($n = 1$; 3.1%).

Table 1

Sociodemographic Characteristics of Participants at Baseline

Baseline characteristics	Full sample ($N = 213$)	Non-clinical ($N = 181$)	Clinical ($N = 32$)	Statistical analyses
	n (%)	*n* (%)	*n* (%)	*p*-value
Age, mean (*SD*)	27.43 (10.81)	25.36 (8.88)	39.09 (13.25)	< .001***
Gender				.015*
Female	165 (77.5)	146 (80.7)	19 (59.4)	
Male	46 (21.6)	34 (18.8)	12 (37.5)	
Divers	2 (0.9)	1 (0.6)	1 (3.1)	
Marital status				.027*
Single	93 (43.7)	75 (41.4)	18 (56.2)	
Married/partnered	119 (55.9)	106 (58.6)	13 (40.6)	
Divorced/widowed	1 (0.5)	0 (0)	1 (3.1)	
Highest level of education				< .001***
Lower secondary school certificate	1 (0.5)	0 (0)	1 (3.1)	
Intermediate secondary school certificate	4 (1.9)	2 (1.1)	2 (6.2)	
University of applied sciences entrance diploma	20 (9.4)	20 (11.1)	0 (0)	
General higher education entrance qualification	106 (50.0)	104 (57.8)	2 (6.2)	
Apprenticeship	25 (11.8)	11 (6.1)	14 (43.8)	
University or postgraduate degree	56 (26.4)	43 (23.9)	13 (40.6)	
Employment status				< .001***
Unemployed	13 (6.1)	0 (0)	13 (40.6)	

Student	154 (72.3)	151 (83.4)	3 (9.4)
Employed	39 (18.3)	24 (13.3)	15 (46.9)
Self-employed	4 (1.9)	4 (2.2)	0 (0)
Retired	1 (0.5)	0 (0)	1 (3.1)
Other	2 (0.9)	2 (1.1)	0 (0)

Note. SD = standard deviation. Age was compared using the Mann-Whitney U test due to non-normal distribution. All categorial variables were compared using Fisher's exact test due to low expected cell counts. *** $p < .001$, * $p < .05$

Table 2

Participants' mean scores of study variables at Baseline

Baseline characteristics	Non-clinical (*N* = 181)	Clinical (*N* = 32)	Statistical analyses	
	Mean (SD)	*Mean (SD)*	*t*-value (211)	*p*-value
ES-G	7.38 (2.03)	3.53 (2.36)	-9.61	< .001***
ES-G Likert	31.83 (7.02)	20.64 (8.16)	-8.11	< .001***
BDI-II	10.45 (10.70)	29.84 (11.04)	9.41	< .001***
WHOQOL-BREF	72.44 (13.10)	52.71 (11.76)	-7.97	< .001***
PWB	83.05 (10.17)	n.a.		
Autonomy	12.50 (2.69)	n.a.		
Environmental mastery	13.57 (2.57)	n.a.		
Personal growth	15.43 (2.30)	n.a.		
Positive relations with others	13.80 (2.94)	n.a.		
Purpose in life	14.01 (2.52)	n.a.		
Self-acceptance	13.75 (2.96)	n.a.		
CD-RISC	25.78 (7.34)	15.56 (7.34)	-7.27	< .001***
GSI	0.64 (0.61)	1.39 (0.69)	6.20	< .001***
WHO-5	59.91 (19.22)	33.62 (17.30)	-7.23	< .001***

Note. SD = standard deviation. ES-G = Euthymia Scale; ES-G Likert = 6-point version of the Euthymia Scale; BDI-II = Beck Depression Inventory – II; WHOQOL-BREF = World Health Organization Quality of Life 21-item version; PWB = Psychological Well-Being Scale was only assessed in the non-clinical sample; CD-RISC = Connor-Davidson Resilience Scale 10-item version; GSI = Global Severity Index of the Brief Symptom Inventory 53-item version. WHO-5 = World Health Organization – 5. All variables were compared using Welch t-tests for unequal variances.

*** p < .001

Correlation analyses

Table 3 presents means, standard deviations, and Spearman rank correlations between the study variables. Spearman correlations were used due to significant deviations from normality in several variables, as indicated by Shapiro-Wilk tests (see Appendix C, Table C1). *p*-values were corrected for multiple comparisons using the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995).

Table 3

Descriptive Statistics and Spearman Correlations among Study Variables

Variable	*n*	*M*	*SD*	1	2	3	4	5
1. ES-G	213	6.80	2.49	—				
2. GSI	213	0.75	0.68	-.70***	—			
3. WHOQOL	213	69.48	14.69	.71***	-.80***	—		
4. CD-RISC	213	24.25	8.18	.64***	-.62***	.64***	—	
5. PWB[a]	181	83.05	10.17	.51***	-.56***	.67***	.62***	—
6. BDI-II	213	13.36	12.78	-.76***	.85***	-.82***	-.61***	-.57***

Note. *n* = number of participants, *M* = mean, *SD* = standard deviation

[a] Psychological Well-Being (PWB) was only assessed in the non-clinical sample.

Euthymia = ES-G, Distress = GSI, Quality of life = WHOQOL-BREF, Resilience = CD-RISC, Depressive Symptoms = BDI-II.

Benjamini Hochberg correction was applied, *** $p < .001$.

Rasch analysis

Rasch analysis revealed some misfit in the conditional item infit statistics (Table 4) with item 4 showing a high item fit (InfitMSQ .10 above the threshold) and Item 5 showing a low item fit (InfitMSQ .10 below the threshold).

Table 4

Conditional Item Fit of Euthymia Scale Items

ES Item	InfitMSQ	Infit thresholds	Infit diff
Item 1	0.95	[0.79, 1.18]	no misfit
Item 2	1.00	[0.85, 1.15]	no misfit
Item 3	1.01	[0.79, 1.20]	no misfit
Item 4	1.42	[0.77, 1.32]	0.10
Item 5	0.89	[0.74, 1.26]	no misfit
Item 6	0.66	[0.77, 1.23]	-0.10
Item 7	0.98	[0.78, 1.22]	no misfit
Item 8	1.04	[0.84, 1.21]	no misfit
Item 9	1.01	[0.80, 1.20]	no misfit
Item 10	1.06	[0.81, 1.33]	no misfit

Note. InfitMSQ = information weighted mean square which is calculated by multiplying the squared standardized residuals by the observed response variance and then divided by the sum of the item response variances. MSQ values are based on conditional calculations (n = 211 complete cases). Thresholds were simulated from a parametric bootstrapping procedure with 200 iterations. Misfit items are highlighted in red.

Principal Component Analysis of Rasch model residuals (PCAR) revealed a first eigenvalue of 1.50 explaining 17.9 % of variance. A parametric bootstrapping procedure with 500 iterations calculated a maximum appropriate cutoff for the first eigenvalue of 1.68 to support unidimensionality.

Residual correlations between item pairs are displayed in Table 5. No correlations above the relative cutoff value of 0.21 were found.

Table 5

Residual Correlations of Euthymia Scale Item Pairs

Item	1	2	3	4	5	6	7	8	9
2	-.03								
3	.16	-.16							
4	-.22	-.03	-.19						
5	-.11	.01	.06	-.18					
6	.01	-.11	.04	-.18	-.12				
7	-.14	-.04	-.17	0	.03	.08			
8	-.24	-.24	-.03	-.09	.01	.03	-.34		
9	-.18	-.18	-.15	-.10	-.15	.01	-.13	.09	
10	-.05	-.10	-.22	-.05	-.27	-.06	-.10	-.06	-.06

Note. Relative cutoff value is 0.205, which is 0.293 above the average correlation (-0.088). Correlations above the relative cutoff are highlighted in red. The relative cut-off value was calculated with a 400 iteration bootstrapping procedure.

To validate the dimensionality assessment of the ES-G, parallel analysis was conducted with 1000 iterations. The analysis suggested that only one factor should be retained. The factor analysis scree plot is shown in Appendix D, Figure D1.

Predictive validity

A predictive modeling approach using nested cross-validation (5 folds, 10 repetitions) was applied to test whether baseline (t0) ES-G total scores or ES-G Likert total scores could predict treatment response. This analysis was conducted on a subsample of $n = 25$ patients who completed both pre- and post-assessment in the clinical trial at the LMU day clinic. A total of 14 patients (56.0%) were classified as responders according to the symptom reduction criterion (≥ 50% decrease in BDI-II total score from t0 to t1), and 15 patients (60.0%) met the well-being criterion (improvement of ≥ 10 points on the WHO-5 from t0 to t1).

The model performance metrics for both criteria and both scale formats are summarized in Table 6.

Sensitivity

The same machine learning approach was used to assess whether baseline ES-G or ES-G Likert total scores could predict group membership (non-clinical participants vs. clinical patients). The full dataset included of 32 clinical patients (15.0%) and 181 non-clinical participants (85.0%). Model performance is also presented in Table 6.

Table 6

Performance of Logistic Regression Models (5-Fold Cross-Validation, 10 Repetitions) in Predicting Treatment Response and Group (clinical vs. non-clinical)

Criterion	n	Predictor	BAC	AUC	TPR	TNR
Response (BDI-II)	25					
~		ES-G	0.61	0.63	0.71	0.48
~		ES-G Likert	0.53	0.60	0.61	0.46
Response (WHO-5)	25					
~		ES-G	0.63	0.69	0.72	0.54
~		ES-G Likert	0.54	0.67	0.58	0.51
Group membership	213					
~		ES-G	0.82	0.88	0.82	0.83
~		ES-G Likert	0.79	0.85	0.78	0.81

Note. Response criteria were defined as: BDI-II = ≥ 50% symptom reduction from t0 to t1; WHO-5 = ≥ 10 points increase from t0 to t1. Group membership was defined

as clinical (day clinic patients) or non-clinical participants. The positive class was defined as "yes" for treatment response and "clinical" for group membership. BAC = balanced accuracy; AUC = area under the ROC curve; TPR = True positive rate (sensitivity); TNR = true negative rate (specificity). ES-G = baseline Euthymia Scale total score; ES-G Likert = baseline Euthymia Scale Likert version total score.

To evaluate the sensitivity to change of the ES-G and it's Likert version, two linear regression models were estimated using ΔBDI-II as the outcome variable. Both models were controlled for centered BDI-II baseline scores.

The model using ΔES-G (centered) as the predictor yielded a significant negative association with symptom change, b = -2.54, $t(22)$ = -6.29, $p < .001$, indicating that greater increase in ES-G was associated with greater symptom reduction. The ES-G change scores explained 64.3% of unique variance in the BDI-II change scores, after accounting for baseline depression, partial R^2 = .64

Similarly, the ES-G Likert version also significantly predicted symptom change, b = -0.64, $t(22)$ = -3.34, p = .003. Partial R^2 of the ES-G Likert was .34.

Assumptions of linear regression were tested for both models. Shapiro-Wilk tests indicated that residuals were normally distributed (p = .812 and p = .716, respectively). Residuals-vs-fitted plots showed no clear patterns, supporting the assumption of homoscedasticity (see Appendix C, Figure C3 and C4)

Clinical Validity

Mean total scores and standard deviations of the ES-G and ES-G Likert version stratified according to the past or current history of MDE are reported in Table 7. Due

to violations of the homogeneity of variance assumption (Appendix C, Table C2), Welch's ANOVA was used. The test revealed a significant effect of group membership on ES-G total scores, $F(4,74.42) = 44.61$, $p < .001$. A bootstrapped estimate of omega squared confirmed a large effect size, $\omega^2 = 0.50$. A Jonckheere–Terpstra trend test with 10,000 permutations revealed a significant decreasing trend in ES-G total scores across the ordered MINI groups, $JT = 3289.5$, $p < .001$.

ES-G Likert total scores also significantly differentiated between the different groups of current or past depression history, Welch's $F(4,81.14) = 28.41$, $p < .001$. Bootstrapped omega squared again indicated a large effect, $\omega^2 = 0.40$. Likewise, a decreasing trend in ES-G Likert total scores was found, $JT = 3662.5$, $p < .001$.

Table 7

Means, Standard Deviations, and Welch's ANOVA of ES-G and ES-G Likert Total Scores stratified by Categories of History of MDE and Current MDE

Scale	Mean (SD)						Welch-ANOVA results	
	Group	0	1	2	3	4		
	Total	Past (-) Current (-)	Past (+) Current (-)	Past (-) Current (±)	Past (+) Current (±)	Past (+) Current (+)		
	N = 206	n = 52	n = 17	n = 52	n = 44	n = 41	*F*-value	ω^2
ES-G	6.88 (2.47)	8.83 (1.31)	7.47 (1.74)	7.60 (1.71)	6.36 (1.66)	3.83 (2.25)	44.61***	.51
ES-G Likert	30.46 (8.07)	36.20 (7.53)	32.70 (4.09)	32.40 (5.97)	28.80 (4.93)	21.50 (6.87)	28.41***	.41

Note. Past (+): total score ≥ 5; Past (-): total score ≤ 4, measured by the Mini International Neuropsychiatric Interview questionnaire for lifetime episode

Current (+): total score ≥ 5; Current (±): 1 ≤ total score ≤ 4; Current (-): score = 0, measured by the Mini International Neuropsychiatric Interview questionnaire for current two weeks episode. ES-G = baseline Euthymia Scale total score; ES-G Likert = baseline Euthymia Scale Likert version total score. Omega squared (ω^2) was estimated using a nonparametric bootstrapping procedure with 1,000 resamples.
*** < .001

Games–Howell post-hoc comparisons revealed that ES-G total scores were significantly lower in all groups with a current or past depressive episode compared to the healthy group (all *p*s < .05), except for the comparison between healthy participants (Group 0) and those in full remission (Group 1), which was not significant ($p = .051$).

No significant differences were found between Group 1 (full remission) and Group 2 (first subthreshold depressive episode; $p = .999$) or between Group 1 and Group 4 (past MDE + current subthreshold symptoms; $p = .188$).

The ES-G Likert total scores showed a similar pattern to the original version. Scores were significantly lower in all clinical groups compared to the healthy group (all *p*s < .05), except for participants in full remission (Group 1), where the difference was not statistically significant ($p = .119$).

No significant difference was found between Group 1 (full remission) and Group 2 (first subthreshold depressive episode; $p = 1.00$).

Mean total scores of the ES-G and the ES-G Likert version stratified by BDI-II symptom severity groups are reported in Table 8. Welch's ANOVA revealed a significant effect of symptom severity on ES-G total scores, $F(3, 49.60) = 78.37$, $p < .001$. A Jonckheere–Terpstra trend test with 10,000 permutations confirmed a significant

decreasing trend in ES-G scores with increasing levels of symptom severity, *JT*= 1592, $p < .001$. A bootstrapped estimate of omega squared confirmed a large effect size, $\omega^2 = .50$.

A similar pattern was observed for the ES-G Likert version. Welch's ANOVA indicated significant mean differences in ES-G Likert total scores between symptom severity groups, $F(3, 54.61) = 57.81$, $p < .001$. A decreasing trend across severity levels was likewise confirmed, $JT = 1714$, $p < .001$. A bootstrapped estimate of omega squared indicated a large effect, $\omega^2 = .45$.

Table 8

Means, Standard Deviations, and Welch's ANOVA of ES-G and ES-G Likert Total Scores stratified by Symptom Severity Groups

Scale	Mean (SD)					Welch-ANOVA results	
	Group Total N = 213	0 minimal n = 132	1 mild n = 28	2 moderate n = 20	3 severe n = 33	*F*-value	ω^2
ES-G	6.80 (2.49)	8.14 (1.46)	5.89 (1.40)	5.20 (2.50)	3.15 (1.91)	78.37***	.50
ES-G Likert	30.15 (8.22)	34.20 (6.18)	28.3 (4.26)	23.8 (6.86)	19.5 (6.11)	57.81***	.45

Note. Symptom severity groups: minimal = total score ≤ 13; mild = 14 ≤ total score ≤ 19; moderate = 20 ≤ total score ≤ 28; severe = 29 ≤ total score, measured by the Beck Depression Inventory-II. ES-G = baseline Euthymia Scale total score; ES-G Likert = baseline Euthymia Scale Likert version total score. Omega squared (ω^2) was estimated using a nonparametric bootstrapping procedure with 1,000 resamples.

*** < .001

A Games–Howell post-hoc test revealed significant group differences between in ES-G scores between all depression severity groups (all *p*s < .05), except for the difference between group 1 (mild depression) and group 2 (moderate depression) (p = .829).

For the ES-G Likert scores, no significant differences were found between the mild and moderate (p = .151), or between the moderate and severe (p = .055) symptom groups. All other group comparisons showed significant differences (all *p*s < .001).

A full display of pairwise comparisons is presented in Appendix E, Tables E1 – E4.

Cutoff Determination

In the non-clinical sample, 53 participants were classified as depressed and 128 as non-depressed based on a BDI-II cutoff of ≥ 13. ROC analysis yielded an AUC of .88. The optimal cutoff score determined by Youden's J was 7.5. For practical purposes, a cutoff of ≥ 7 is recommended, which resulted in a sensitivity of 88.7% and specificity of 71.9% (balanced accuracy (BAC) = 80.3%). The ROC curve is shown in Appendix F, Figure F1.

Among day-clinic patients, 27 individuals met the criterion for depression, while 5 did not, based on a BDI-II cutoff of ≥ 19. The ROC analysis produced an AUC of .94. Youden's J indicated an optimal cutoff of 4.5. For the clinical use, a cutoff of ≥ 5 seems to be appropriate. At this value, sensitivity was 92.6% and specificity was 80.0%, resulting in a balanced accuracy (BAC) of 86.3%. The ROC curve is provided in Appendix F, Figure F2.

Incremental Validity

To assess the incremental validity of the ES-G, two hierarchical regression models were conducted for each ES-G version (dichotomous vs. Likert), predicting psychological well-being (PWB total score). Both models were controlled for sex, age, and education. In the first model, adding the ES-G at step 2 led to a significant increase in explained variance, $\Delta R^2 = .08$, $F(1,167) = 21.32$, $p < .001$. In the second model, the ES-G Likert version also accounted for a significant increase, $\Delta R^2 = .06$, $F(1,167) = 16.82$, $p < .001$.

Residual diagnostics showed no substantial deviations from homoscedasticity (Appendix C, Figures C5 and C6). Although Shapiro-Wilk tests of the standardized residuals indicated deviations from normality (p's < .001), linear regression and model comparison tests are generally robust to such violations in large samples (e.g., (Lumley et al., 2002). Given that the primary interest was in changes in explained variance (ΔR^2), and no serious violations of other assumptions were observed, the results are considered reliable.

Comparison of the Self-Adapted 6-Point Likert Version of the ES-G with the Original Version

A comparison of the original and Likert versions of the ES-G across all validation objectives is presented in Table 9.

Table 9

Comparison of the Original and Likert Versions of the ES-G Across Validation Objectives

Analysis	Metric	ES-G Original	ES-G Likert
Predictive modeling	BAC		

- Treatment response (BDI-II)		.61	.53
- Treatment response (WHO-5)		.63	.54
- Predicting group membership (clinical vs. non-clinical)		.82	.79
Linear regression	partial R^2 / ΔR^2		
- Sensitivity to change		.64	.34
- Incremental validity		.08	.06
Welch's ANOVA	ω^2		
- Depression history (MINI)		.51	.41
- Symptom severity (BDI-II)		.50	.45

Note. BDI-II = Beck Depression Inventory-II; WHO-5 = World Health Organization – 5; MINI = Mini International Neuropsychiatric Interview questionnaire, ES-G = Euthymia Scale; ES-G Likert = 6-point version of the Euthymia Scale; BAC = balanced accuracy; partial R^2 = proportion of variance uniquely explained by ES-G change scores; ΔR^2 = increase in explained variance; ω^2 = estimated proportion of explained variance

Discussion

Summary of Main Findings

The aim of this study was to validate the German version of the Euthymia Scale (ES-G) through a comprehensive clinimetric analysis. Following the CLIPROM criteria for patient-reported outcome measures (Carrozzino, Patierno, et al., 2021), this study assessed concurrent validity, dimensionality, predictive validity, sensitivity, clinical validity, and incremental validity of the ES-G, employing state-of-the-art statistical methods. Additionally, a statistically valid cutoff score for the ES-G as a screener for depression was determined and a self-adapted 6-point Likert version of the ES-G was compared with the original version.

The Euthymia Scale (ES-G) demonstrated good concurrent validity, correlating positively with measures of positive mental health (quality of life, resilience, psychological well-being) and negatively with measures of psychological distress. Although Pearson correlations were preregistered, we observed significant violations of normality in several variables. Therefore, Spearman rank correlations were computed instead. The results did not differ in terms of direction or significance and are thus interpreted in line with the original hypotheses. All correlations were highly significant and in the hypothesized directions. Effect sizes were all high according to benchmarks proposed by Cohen (1988), with the strongest association observed between the ES-G and depressive symptoms ($\rho = -.76$) and quality of life ($\rho = .71$). These findings support the convergent validity of the ES-G and are in line with the theoretical framework (Guidi & Fava, 2022) of what the Euthymia Scale (ES) should measure; that is a lack of mood disturbances, the presence of positive affect, and integration as defined by Jahoda (1959).

Several indicators of dimensionality of the ES-G were assessed using Rasch measurement analysis for dichotomous data. Conditional item fit statistics revealed some misfit. Item 4 displayed a high infit value (underfit), and item 6 showed a low infit value (overfit). Overfit indicates that responses may be too predictable and provide little information and is generally not considered an indicator for multidimensionality. Underfit may be an indicator of multidimensionality (Johansson, 2025b). Therefore Item 4 underfitting the Rasch model was of potential concern regarding the dimensionality. However, follow-up dimensionality tests did not support this concern. Principal component analysis of Rasch model residuals (PCAR) and analysis of residual correlations of item pairs showed no signs of multidimensionality. The first eigenvalue was below the estimated highest first eigenvalue and all residual correlations

were below the bootstrapped relative cutoff. Furthermore, a parallel analysis suggested a unidimensional structure. While these findings suggest a unidimensional structure in this sample, previous research has identified two underlying dimensions (Carrozzino, Christensen, Mansueto, et al., 2021; Carrozzino et al., 2019).

The baseline total scores the ES-G, whether in the dichotomous or the 6-point Likert adaption, showed limited ability in predicting treatment response. AUC values were barely above chance level (AUCs = .63 and .60 for BDI-II; .69 and .67 for WHO-5). This indicates that ES-G total score alone may not serve as a reliable predictor of psychotherapy outcome. However, this finding must be interpreted with caution due to the small sample size of this analysis (n = 25).

Significantly better results were observed when predicting group membership (clinical vs. non-clinical). The original ES-G achieved an AUC of .88 and a BAC of .82, while the Likert version showed slightly lower values (AUC = .85, BAC = .79). These results indicate a strong ability of the ES-G to discriminate between clinical and non-clinical populations which is considered an important criterion for clinimetric outcome measures (Giovanni A. Fava et al., 2018).

Both versions of the ES-G demonstrated good sensitivity to change, as reflected by their significant associations with symptom improvement. This supports their ability to reflect outcome changes in clinical trials, which is a key requirement for clinimetric instruments (Kellner, 1972).

Findings of the Welch's ANOVAs showed that the ES-G is a highly sensitive clinimetric index, effectively differentiating between past and current depression history groups. It distinguished particularly well between healthy participants, that have never experienced a depressive episode (MDE), and all other subthreshold or full MDE

groups, and perhaps most importantly, between individuals with current MDE and those with subthreshold symptoms. This ability to detect residual or subclinical symptoms following remission is especially relevant in the clinical evaluation of recovery, as such symptoms are known to prevent full remission and substantially increase the risk of relapse (Conradi et al., 2008, 2012; Verhoeven et al., 2018). However, ES-G total scores did not significantly differ between individuals in full remission and those with current subthreshold symptoms, indicating limited discriminative ability within the subthreshold range. These findings were supported from a second ANOVA, distinguishing ES-G total scores by BDI-II symptom groups: while participants with no or minimal depressive symptoms (BDI-II ≤ 13) showed significantly higher ES-G scores than all other symptom severity groups, no significant difference was observed between groups with mild and moderate depressive symptoms. These findings are in line with previous results reported for the Japanese version of the Euthymia Scale (Sasaki et al., 2021).

Results from the ROC analyses suggest, that the ES-G total score may serve as a useful screening indicator for depression in both clinical and non-clinical populations. The ES-G demonstrated good discriminative ability in the non-clinical sample (AUC = .88) and excellent performance in the clinical sample (AUC = .94). In the non-clinical sample, a cutoff score of ≥ 7 provided a sensitivity of 88.7% and specificity of 71.9% (BAC = .80). In the clinical sample, a slightly lower threshold of ≥ 5 yielded the best balance (BAC = .86) between sensitivity (92.6%) and specificity (80.0%). These group-specific cutoffs are consistent with recommendations by von Glischinski et al. (2019), which propose different BDI-II cutoffs depending on clinical setting. However, even cutoff points identified as optimal would miss certain individuals with

depression and wrongfully classify others. This highlights the importance of clinical interviews when diagnosing depression.

Both versions of the ES-G also made a meaningful incremental contribution to the prediction of psychological well-being (PWB), beyond what was explained by the WHO-5.

Comparing the original dichotomous version of the ES-G with the adapted Likert version a clear pattern was found: the original version constantly outperformed the Likert version across all validation metrics, including predictive accuracy (BAC), sensitivity to change (R^2), group differentiation (ω^2), and incremental validity (ΔR^2). These findings support the assumption made by the original authors that the dichotomous response format increases the sensitivity of the scale (Giovanni A. Fava & Bech, 2016; Guidi & Fava, 2022).

Objective 9 entfern (obmitted)t. Redundant. Changed order of objectives.

Keine balancierten Gruppen. Nicht representative, selecting bias (uni)

Anderer cut off für Depression

Generalizability limited

Schwächen:

A parametric bootstrap function has been implemented in easyRasch to determine a potentially appropriate cutoff value for the largest PCAR eigenvalue, but it has not been systematically evaluated yet. Below is an example, illustrated with a histogram of the simulated distribution of largest eigenvalues, the 99th percentile and the max value. If the bootstrap turns out to provide an appropriate cutoff value, it still needs to be used together with checking item fit (or item-restscore) and residual correlations (local dependence) to evaluate unidimensionality.

Although the preregistration included testing incremental validity across PWB subscales, the present report focuses on the total score to preserve brevity and clarity.

While both versions of the ES-G showed significant incremental validity beyond the WHO-5, the original scale accounted for a slightly larger proportion of additional variance in psychological well-being ($\Delta R^2 = .08$) compared to the Likert version ($\Delta R^2 = .06$). This suggests that [...insert interpretation based on theory or measurement format...].

Across all criteria, the original version of the ES-G showed marginally superior psychometric performance.

Implications

Strengths and Limitations

ES does not include PWB, CIE does.

MINI as self report

In order to measure Euthymia as defined you need several rating scales (distress, ES, Kellner Symptom's Questionnaire)

Future Research

Conclusion

References

American Psychiatric Association, A., Association, A. P., & Others. (1994). *Diagnostic and statistical manual of mental disorders: DSM-IV* (Vol. 4). American psychiatric association Washington, DC.

Angermeyer, M. C., Kilian, R., & Matschinger, H. (2000). World health organization quality of life (WHOQOL). *Göttingen: Hogrefe*.

Apgar, V. (1953). A proposal for a new method of evaluation of the newborn infant. *Current Researches in Anesthesia & Analgesia*, *32*(4), 260–267. https://doi.org/10.1213/00000539-195301000-00041

Bagby, R. M., Taylor, G. J., & Parker, J. D. (1994). The Twenty-item Toronto Alexithymia Scale--II. Convergent, discriminant, and concurrent validity. *Journal of Psychosomatic Research*, *38*(1), 33–40. https://doi.org/10.1016/0022-3999(94)90006-x

Bech, P. (2004). Modern psychometrics in clinimetrics. *Psychotherapy and Psychosomatics*, *73*(3), 134–138. https://www.jstor.org/stable/48510813

Bech, P. (2012). *Clinical psychometrics*. https://books.google.com/books?hl=en&lr=&id=pyNw_eDw5kMC&oi=fnd&pg=PT10&dq=Bech+P.+Clinical+psychometrics.+Oxford:+Wiley-+Blackwe&ots=5h282xQ37t&sig=JjzeZnOocPIfVLurruXsn49bhbI

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *BDI-II, Beck Depression Inventory: Manual*. Psychological Corporation.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, *57*(1), 289–300.

Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., Casalicchio, G., & Jones, Z. M. (2016). mlr: Machine Learning in R. *Journal of Machine Learning Research: JMLR*, *17*(170), 170:1-170:5. http://www.jmlr.org/papers/v17/15-066.html

Blanchflower, D. G., & Oswald, A. J. (2011). International happiness: A new view on the measure of performance. *The Academy of Management Perspectives*, *25*(1), 6–22. https://doi.org/10.5465/amp.25.1.6

Buecker, S., Luhmann, M., Haehner, P., Bühler, J. L., Dapp, L. C., Luciano, E. C., & Orth, U. (2023). The development of subjective well-being across the life span: A meta-analytic review of longitudinal studies. *Psychological Bulletin*, *149*(7–8), 418–446. https://doi.org/10.1037/bul0000401

Campbell-Sills, L., & Stein, M. B. (2007). Psychometric analysis and refinement of the Connor-davidson Resilience Scale (CD-RISC): Validation of a 10-item measure of resilience. *Journal of Traumatic Stress*, *20*(6), 1019–1028. https://doi.org/10.1002/jts.20271

Carrozzino, D. (2019). Clinimetric approach to rating scales for the assessment of apathy in Parkinson's disease: A systematic review. *Progress in Neuro-Psychopharmacology & Biological Psychiatry*, *94*(109641), 109641. https://doi.org/10.1016/j.pnpbp.2019.109641

Carrozzino, D., Christensen, K. S., & Cosci, F. (2021). Construct and criterion validity of patient-reported outcomes (PROs) for depression: A clinimetric comparison. *Journal of Affective Disorders*, *283*, 30–35. https://doi.org/10.1016/j.jad.2021.01.043

Carrozzino, D., Christensen, K. S., Mansueto, G., Brailovskaia, J., Margraf, J., & Cosci, F. (2021). A clinimetric analysis of the euthymia, resilience, and positive mental health scales. *Journal of Affective Disorders*, *294*, 71–76. https://doi.org/10.1016/j.jad.2021.07.001

Carrozzino, D., Patierno, C., Guidi, J., Berrocal Montiel, C., Cao, J., Charlson, M. E., Christensen, K. S., Concato, J., De Las Cuevas, C., de Leon, J., Eöry, A., Fleck, M. P., Furukawa, T. A., Horwitz, R. I., Nierenberg, A. A., Rafanelli, C., Wang, H., Wise, T. N., Sonino, N., & Fava, G. A. (2021). Clinimetric Criteria for Patient-Reported Outcome Measures. *Psychotherapy and Psychosomatics*, *90*(4), 222–232. https://doi.org/10.1159/000516599

Carrozzino, D., Svicher, A., Patierno, C., Berrocal, C., & Cosci, F. (2019). The Euthymia Scale: A Clinimetric Analysis [Review of *The Euthymia Scale: A Clinimetric Analysis*]. *Psychotherapy and Psychosomatics*, *88*(2), 119–121. https://doi.org/10.1159/000496230

Cella, D., Yount, S., Rothrock, N., Gershon, R., Cook, K., Reeve, B., Ader, D., Fries, J., Bruce, B., & Rose, M. (2007). The patient-Reported Outcomes Measurement Information System (PROMIS): Progress of an NIH Roadmap cooperative group during its first two years. *Medical Care*, *45*, S3–S11. https://doi.org/10.1097/01.mlr.0000258615.42478.55

Cella, David, Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., Amtmann, D., Bode, R., Buysse, D., Choi, S., Cook, K., Devellis, R., DeWalt, D., Fries,

J. F., Gershon, R., Hahn, E. A., Lai, J.-S., Pilkonis, P., Revicki, D., … PROMIS Cooperative Group. (2010). The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005-2008. *Journal of Clinical Epidemiology*, *63*(11), 1179–1194. https://doi.org/10.1016/j.jclinepi.2010.04.011

Charter, R. A. (1999). Sample size requirements for precise estimates of reliability, generalizability, and validity coefficients. *Journal of Clinical and Experimental Neuropsychology*, *21*(4), 559–566. https://doi.org/10.1076/jcen.21.4.559.889

Chou, Y.-T., & Wang, W.-C. (2010). Checking dimensionality in item response models with principal component analysis on standardized residuals. *Educational and Psychological Measurement*, *70*(5), 717–731. https://doi.org/10.1177/0013164410379322

Christensen, K. B., Makransky, G., & Horton, M. (2017). Critical values for yen's Q3: Identification of local dependence in the Rasch model using residual correlations. *Applied Psychological Measurement*, *41*(3), 178–194. https://doi.org/10.1177/0146621616677520

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

Connor, K. M., & Davidson, J. R. T. (2003). Development of a new resilience scale: the Connor-Davidson Resilience Scale (CD-RISC). *Depression and Anxiety*, *18*(2), 76–82. https://doi.org/10.1002/da.10113

Conradi, H. J., de Jonge, P., & Ormel, J. (2008). Prediction of the three-year course of recurrent depression in primary care patients: different risk factors for different outcomes. *Journal of Affective Disorders*, *105*(1–3), 267–271. https://doi.org/10.1016/j.jad.2007.04.017

Conradi, H. J., Ormel, J., & de Jonge, P. (2012). Symptom profiles of DSM-IV-defined remission, recovery, relapse, and recurrence of depression: the role of the core symptoms: Research article: Symptom profiles of remissions and recoveries. *Depression and Anxiety*, *29*(7), 638–645. https://doi.org/10.1002/da.21960

Cronbach, L., & Meehl, P. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/H0040957

Deci, E. L., & Ryan, R. M. (2008). Hedonia, eudaimonia, and well-being: an introduction. *Journal of Happiness Studies*, *9*(1), 1–11. https://doi.org/10.1007/s10902-006-9018-1

Derogatis, L. R. (1993). *BSI, Brief Symptom Inventory: Administration, Scoring & Procedures Manual*. National Computer Systems. https://play.google.com/store/books/details?id=9JTFDAEACAAJ

Diener, E. (1984). Subjective well-being. *Psychological Bulletin*, *95*(3), 542–575. https://doi.org/10.1037/0033-2909.95.3.542

Diener, E., Wirtz, D., Tov, W., Kim-Prieto, C., Choi, D.-W., Oishi, S., & Biswas-Diener, R. (2010). New well-being measures: Short scales to assess flourishing and positive and negative feelings. *Social Indicators Research*, *97*(2), 143–156. https://doi.org/10.1007/s11205-009-9493-y

Dodge, R., Daly, A. P., Huyton, J., & Sanders, L. (2012). The challenge of defining wellbeing. *International Journal of Wellbeing*, *2*(3), 222–235. https://doi.org/10.5502/IJW.V2.I3.4

Endermann, M. (2005). The Brief Symptom Inventory (BSI) as a screening tool for psychological disorders in patients with epilepsy and mild intellectual disabilities in residential care. *Epilepsy & Behavior: E&B*, *7*(1), 85–94. https://doi.org/10.1016/j.yebeh.2005.03.018

Fava, G. A. (2016). *Well-Being Therapy: Treatment Manual and Clinical Applications*. Karger Medical and Scientific Publishers. https://play.google.com/store/books/details?id=hPmfCwAAQBAJ

Fava, G. A., Tomba, E., & Sonino, N. (2012). Clinimetrics: the science of clinical measurements. *International Journal of Clinical Practice*, *66*(1), 11–15. https://doi.org/10.1111/j.1742-1241.2011.02825.x

Fava, Giovanni A., & Bech, P. (2016). The Concept of Euthymia. *Psychotherapy and Psychosomatics*, *85*(1), 1–5. https://doi.org/10.1159/000441244

Fava, Giovanni A., & Belaise, C. (2005). A discussion on the role of clinimetrics and the misleading effects of psychometric theory. *Journal of Clinical Epidemiology*, *58*(8), 753–756. https://doi.org/10.1016/j.jclinepi.2004.12.006

Fava, Giovanni A., Carrozzino, D., Lindberg, L., & Tomba, E. (2018). The clinimetric approach to psychological assessment: A tribute to per Bech, MD (1942-2018). *Psychotherapy and Psychosomatics*, *87*(6), 321–326. https://doi.org/10.1159/000493746

Fava, Giovanni A., & Guidi, J. (2020a). The pursuit of euthymia. *World Psychiatry: Official Journal of the World Psychiatric Association* , *19*(1), 40–50. https://doi.org/10.1002/wps.20698

Fava, Giovanni A., & Guidi, J. (2020b). Das Streben nach Euthymie. *Ärztliche Psychotherapie Und Psychosomatische Medizin*, *15*(3), 149–165. https://doi.org/10.21706/aep-15-3-149

Fava, Giovanni A., Ruini, C., & Rafanelli, C. (2004). Psychometric theory is an obstacle to the progress of clinical research. *Psychotherapy and Psychosomatics*, *73*(3), 145–148. https://doi.org/10.1159/000076451

Feinstein, A. (1987). Clinimetric perspectives. *Journal of Chronic Diseases*, *40*(6), 635–640. https://doi.org/10.1016/0021-9681(87)90027-0

Feinstein, A. R. (1983). An additional basic science for clinical medicine: IV. The development of clinimetrics. *Annals of Internal Medicine*, *99*(6), 843–848. https://doi.org/10.7326/0003-4819-99-6-843

Feinstein, Alvan R. (1987). Clinimetrics. *Yale University Press*.

Franke, G. H., & Derogatis, L. R. (2000). *BSI: brief sympton inventory von LR Derogatis; Kurzform der SCL-90-R); deutsche Version*. https://scholar.google.com/citations?user=rtdHW9AAAAAJ&hl=en&oi=sra

Frost, M. H., Reeve, B. B., Liepa, A. M., Stauffer, J. W., Hays, R. D., & Mayo/FDA Patient-Reported Outcomes Consensus Meeting Group; (2007). What is sufficient evidence for the reliability and validity of patient-reported outcome measures? *Value in Health: The Journal of the International Society for Pharmacoeconomics and Outcomes Research*, *10 Suppl 2*, S94–S105. https://doi.org/10.1111/j.1524-4733.2007.00272.x

Group, W., & Others. (1998). Development of the World Health Organization WHOQOL-BREF quality of life assessment. *Psychological Medicine*, *28*(3), 551–558.

Guidi, J., & Fava, G. A. (2022). The Clinical Science of Euthymia: A Conceptual Map. *Psychotherapy and Psychosomatics*, *91*(3), 156–167. https://doi.org/10.1159/000524279

Hautzinger, M., Keller, F., & Kühner, C. (2006). *Beck depressions-inventar (BDI-II)*. Harcourt Test Services.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. https://doi.org/10.1109/tkde.2008.239

Health Organization, W. (1998). Wellbeing measures in primary health care/the depcare project. *Copenhagen: WHO Regional Office for Europe*.

Hicks, S., Tinkler, L., & Allin, P. (2013). Measuring subjective well-being and its potential role in policy: Perspectives from the UK office for national statistics. *Social Indicators Research*, *114*(1), 73–86. https://doi.org/10.1007/s11205-013-0384-x

Irwin, T. (2019). *Nicomachean ethics*. Hackett Publishing. https://books.google.com/books?hl=en&lr=&id=TSusDwAAQBAJ&oi=fnd&pg=PP1&dq=Aristotle,+C.,+trans.+Terence+Irwin+&ots=65bCk9E9Ee&sig=NufoiUdNysbrmdIWyRtLlZZYrMU

Jahoda, M. (1959). Current concepts of positive mental health. *The American Journal of the Medical Sciences*, *238*, 527. https://doi.org/10.1037/11258-000

Johansson, M. (2025a). *easyRasch: Psychometric Analysis in R with Rasch Measurement Theory*. https://github.com/pgmj/easyRasch

Johansson, M. (2025b). Detecting item misfit in Rasch models. *Educational Methods and Psychometrics*, *3*(2025), 1–58. https://doi.org/10.61186/emp.2025.5

Johansson, M., Preuter, M., Karlsson, S., Möllerberg, M.-L., Svensson, H., & Melin, J. (2023). *Valid and reliable? Basic and expanded recommendations for psychometric reporting and quality assessment*. https://osf.io/preprints/3htzc/

Jones, T. D., & Feinstein, A. R. (1982). T. Duckett Jones Memorial Lecture. The Jones criteria and the challenges of clinimetrics. *Circulation*, *66*(1), 1–5. https://doi.org/10.1161/01.CIR.66.1.1

Kellner, R. (1972). 2. Improvement criteria in drug trials with neurotic patients. *Psychological Medicine*, *2*(1), 73–80. https://doi.org/10.1017/s0033291700045645

Kim, D., De Ayala, R. J., Ferdous, A. A., & Nering, M. L. (2011). The comparative performance of conditional independence indices. *Applied Psychological Measurement*, *35*(6), 447–471. https://doi.org/10.1177/0146621611407909

Kühner, C., Bürger, C., Keller, F., & Hautzinger, M. (2007). Reliabilität und Validität des revidierten Beck-Depressionsinventars (BDI-II). *Der Nervenarzt*, *78*, 651–656. https://doi.org/10.1007/s00115-006-2098-7

Lumley, T., Diehr, P., Emerson, S., & Chen, L. (2002). The importance of the normality assumption in large public health data sets. *Annual Review of Public Health*, *23*(1), 151–169. https://doi.org/10.1146/annurev.publhealth.23.100901.140546

Metz, C. E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, *8*(4), 283–298. https://doi.org/10.1016/s0001-2998(78)80014-2

Mokken, R. J. (1970). *A theory and procedure of scale analysis: with applications in political research*. https://library.wur.nl/WebQuery/titel/411763

Mokkink, L. B., de Vet, H. C. W., Prinsen, C. A. C., Patrick, D. L., Alonso, J., Bouter, L. M., & Terwee, C. B. (2018). COSMIN Risk of Bias checklist for

systematic reviews of Patient-Reported Outcome Measures. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation*, *27*(5), 1171–1179. https://doi.org/10.1007/s11136-017-1765-4

Mokkink, L. B., Terwee, C. B., Knol, D. L., Stratford, P. W., Alonso, J., Patrick, D. L., Bouter, L. M., & de Vet, H. C. W. (2006). Protocol of the COSMIN study: COnsensus-based Standards for the selection of health Measurement INstruments. *BMC Medical Research Methodology*, *6*(1), 2. https://doi.org/10.1186/1471-2288-6-2

Mokkink, Lidwine B., Prinsen, C. A. C., Bouter, L. M., Vet, H. C. W. de, & Terwee, C. B. (2016). The COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) and how to select an outcome measurement instrument. *Brazilian Journal of Physical Therapy*, *20*(2), 105–113. https://doi.org/10.1590/bjpt-rbf.2014.0143

Mokkink, Lidwine B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., Bouter, L. M., & de Vet, H. C. W. (2010). The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation*, *19*(4), 539–549. https://doi.org/10.1007/s11136-010-9606-8

Müller, M. (2020). Item fit statistics for Rasch analysis: can we trust them? *Journal of Statistical Distributions and Applications*, *7*(1). https://doi.org/10.1186/s40488-020-00108-7

Naci, H., & Ioannidis, J. (2015). Evaluation of wellness determinants and interventions by citizen scientists. *JAMA*, *314*(2), 121–122. https://doi.org/10.1001/jama.2015.6160

Oishi, S., & Tay, L. (2019). Gender differences in subjective well-being. *Handbook of Well-Being*. https://www.researchgate.net/profile/Louis-Tay/publication/375083911_Handbook_of_Wellbeing/links/653fd5183cc79d48c5bc41ac/Handbook-of-Wellbeing.pdf#page=359

Organization, W. H., & Others. (2021). Comprehensive mental health action plan 2013--2030. In *Comprehensive mental health action plan 2013--2030*. pesquisa.bvsalud.org. https://pesquisa.bvsalud.org/portal/resource/pt/who-345301

R Core Team. (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rasch, G. (1993). *Probabilistic Models for Some Intelligence and Attainment Tests*. MESA Press, 5835 S. Kimbark Ave., Chicago, IL 60637; e-mail: MESA@uchicago.edu; web address: www.rasch.org; telephone: 773-702-1596 fax: 773-834-0326 ($20). https://eric.ed.gov/?id=ED419814

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. In *BMC Bioinformatics* (Vol. 12, p. 77).

Rothrock, N. E., Kaiser, K. A., & Cella, D. (2011). Developing a valid patient-reported outcome measure. *Clinical Pharmacology and Therapeutics*, *90*(5), 737–742. https://doi.org/10.1038/clpt.2011.195

Ryff, C. D., & Keyes, C. L. M. (1995). The structure of psychological well-being revisited. *Journal of Personality and Social Psychology*, *69*(4), 719–727. https://doi.org/10.1037//0022-3514.69.4.719

Ryff, Carol D. (1989). Happiness is everything, or is it? Explorations on the meaning of psychological well-being. *Journal of Personality and Social Psychology*, *57*(6), 1069–1081. https://doi.org/10.1037/0022-3514.57.6.1069

Sarubin, N., Gutt, D., Giegling, I., Bühner, M., Hilbert, S., Krähenmann, O., Wolf, M., Jobst, A., Sabaß, L., Rujescu, D., Falkai, P., & Padberg, F. (2015). Erste Analyse der psychometrischen Eigenschaften und Struktur der deutschsprachigen 10- und 25-Item Version der Connor-Davidson Resilience Scale (CD-RISC). *Zeitschrift Für Gesundheitspsychologie*, *23*(3), 112–122. https://doi.org/10.1026/0943-8149/a000142

Sasaki, N., Carrozzino, D., & Nishi, D. (2021). Sensitivity and concurrent validity of the Japanese version of the Euthymia scale: a clinimetric analysis. *BMC Psychiatry*, *21*(1), 482. https://doi.org/10.1186/s12888-021-03494-7

Sasaki, N., & Nishi, D. (2022). *Euthymia scale as a predictor of depressive symptoms: a one-year follow-up longitudinal study*.

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*, *47*(5), 609–612. https://doi.org/10.1016/j.jrp.2013.05.009

Sechrest, L. (1963). Incremental validity : A recommendation. *Educational and Psychological Measurement*, *23*(1), 153–158. https://doi.org/10.1177/001316446302300113

Sheehan, D. V., Lecrubier, Y., Sheehan, K. H., Amorim, P., Janavs, J., Weiller, E., Hergueta, T., Baker, R., Dunbar, G. C., & Others. (1998). The Mini-International Neuropsychiatric Interview (MINI): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *The Journal of Clinical Psychiatry*, *59*(20), 22–33.

Smith, E. V., Jr. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement*, *3*(2), 205–231. https://www.ncbi.nlm.nih.gov/pubmed/12011501

Strauss, M. E., & Smith, G. T. (2009). Construct validity: advances in theory and methodology. *Annual Review of Clinical Psychology*, *5*(1), 1–25. https://doi.org/10.1146/annurev.clinpsy.032408.153639

Tomba, E., & Bech, P. (2012). Clinimetrics and clinical psychometrics: macro- and micro-analysis. *Psychotherapy and Psychosomatics*, *81*(6), 333–343. https://doi.org/10.1159/000341757

Topp, C. W., Østergaard, S. D., Søndergaard, S., & Bech, P. (2015). The WHO-5 Well-Being Index: a systematic review of the literature. *Psychotherapy and Psychosomatics*, *84*(3), 167–176.

Verhoeven, F. E. A., Wardenaar, K. J., Ruhé, H. G. E., Conradi, H. J., & de Jonge, P. (2018). Seeing the signs: Using the course of residual depressive symptomatology to predict patterns of relapse and recurrence of major depressive disorder. *Depression and Anxiety*, *35*(2), 148–159. https://doi.org/10.1002/da.22695

von Glischinski, M., von Brachel, R., & Hirschfeld, G. (2019). How depressed is "depressed"? A systematic review and diagnostic meta-analysis of optimal

cut points for the Beck Depression Inventory revised (BDI-II). *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation*, *28*(5), 1111–1118. https://doi.org/10.1007/s11136-018-2050-x

Wood, A. M., & Tarrier, N. (2010). Positive Clinical Psychology: a new vision and strategy for integrated research and practice. *Clinical Psychology Review*, *30*(7), 819–829. https://doi.org/10.1016/j.cpr.2010.06.003

Wood, W., Rhodes, N., & Whelan, M. (1989). Sex differences in positive well-being: A consideration of emotional style and marital status. *Psychological Bulletin*, *106*(2), 249–264. https://doi.org/10.1037/0033-2909.106.2.249

Wright, J. G., & Feinstein, A. R. (1992). A comparative contrast of clinimetric and psychometric methods for constructing indexes and rating scales. *Journal of Clinical Epidemiology*, *45*(11), 1201–1218. https://doi.org/10.1016/0895-4356(92)90161-f

Zweig, M. H., & Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, *39*(4), 561–577. https://www.ncbi.nlm.nih.gov/pubmed/8472349

Appendix

Appendix A – German Adaption of the Euthymia Scale

Table A1

English and German items of the Euthymia Scale (ES)

Item	English version	German version	Answer format
1	If I become sad, anxious or angry it is for a short time	Wenn ich traurig, ängstlich oder wütend werde, hält es nur für kurze Zeit an	richtig/falsch
2	I do not keep thinking about negative experiences	Ich denke nicht ständig über negative Erfahrungen nach	richtig/falsch
3	I am able to adapt to changing situations	Ich kann mich an veränderte Situationen anpassen	richtig/falsch
4	I try to be consistent in my attitudes and behaviors	Ich bemühe mich um beständige Einstellungen und Verhaltensweisen	richtig/falsch
5	Most of the time I can handle stress	Meistens bin ich in der Lage, mit Stress gut umzugehen	richtig/falsch
6	I generally feel cheerful and in good spirits	Ich bin im Allgemeinen froh und guter Laune	richtig/falsch
7	I generally feel calm and relaxed	Ich bin im Allgemeinen ruhig und entspannt	richtig/falsch
8	I generally feel active and vigorous	Ich bin im Allgemeinen aktiv und energisch	richtig/falsch
9	My daily life is filled with things that interest me	Mein Alltagsleben ist voller Dinge, die mich interessieren	richtig/falsch
10	I wake up feeling fresh and rested	Ich fühle mich beim Aufwachen frisch und ausgeruht	richtig/falsch

Appendix B – Grouping Criteria

Table B1

Grouping strategy for depression history

Group	Past MDE	Current MDE	Interpretation
0	no	no	No history of MDE - healthy
1	yes	no	Full remission
2	no	subthreshold	First subthreshold episode
3	yes	subthreshold	History of MDE + current subthreshold
4	yes	yes	History of MDE + current MDE

Note. Past MDE: endorsed ≥ 5 symptoms = yes, < 5 symptoms = no, based on the MINI questionnaire for lifetime episode; Current MDE: 5−9 symptoms = yes, 1−4 symptoms = subthreshold, 0 symptoms = no, based on the MINI questionnaire for current 2-week episodes.

Appendix C – Assumption Checks for Parametric Analyses

Table C1

Kurtosis, Skew and Shapiro Wilk Normality-Test Results of Study Variables

Variable	Skew	Kurtosis	Shapiro wilk test *p*-values
ES-G	-0.75	-0.19	< .001***
ES-G Likert	-0.34	0.24	.033*
GSI	1.05	0.30	< .001***
WHOQOL-BREF	-0.53	-0.43	< .001***
CDRISC	-0.47	-0.02	< .001***
PWB	-0.50	-0.08	.002**
BDI-II	1.16	0.71	< .001***
WHO-5	-0.34	-0.56	< .001***

Note. *p*-values < .05 indicate deviation from normality

Table C2

Levene's and Shapiro-Wilk Tests for Welch ANOVA Models

ANOVA Model	Scale / Group	Levene's test *p*-values	Shapiro-Wilk test *p*-values
MINI history of MDE and current MDE Groups	ES-G	< .001***	
	0		< .001***
	1		.135
	2		.004**
	3		< .001***
	4		.028*
MINI history of MDE and current MDE Groups	ES-G Likert	.024*	
	0		.088
	1		.969
	2		.218
	3		.370
	4		.451
BDI-II symptom Severity Groups	ES-G	< .001***	
	0		< .001***
	1		.254
	2		.399
	3		.006**
BDI-II symptom Severity Groups	ES-G Likert	.117	
	0		.012**
	1		.679
	2		.610
	3		.134

Note. Levene's test *p*-values < .05 indicate violation of homogeneity of variances; Shapiro-Wil test *p*-values < .05 indicate deviation from normality

Figure C3

Residuals vs. Fitted Plot for ES-G Model from Sensitivity to Change Analysis

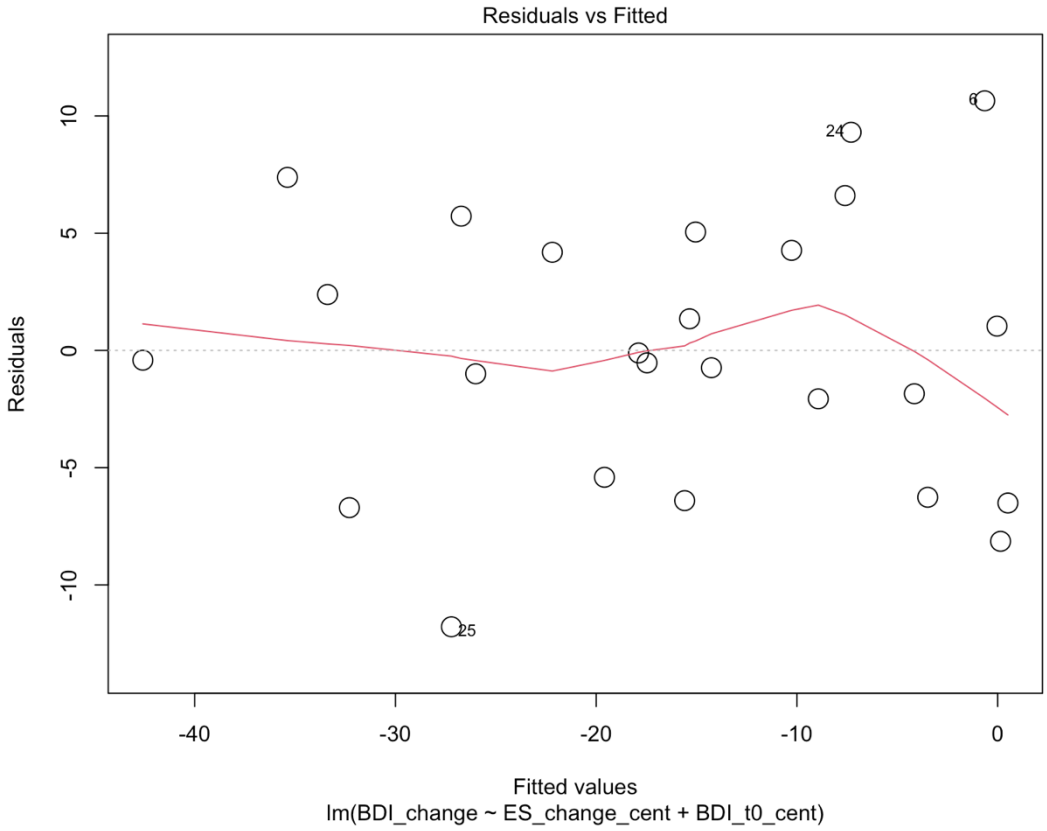


Figure C4

Residuals vs. Fitted Plot for ES-G Likert Model from Sensitivity to Change Analysis

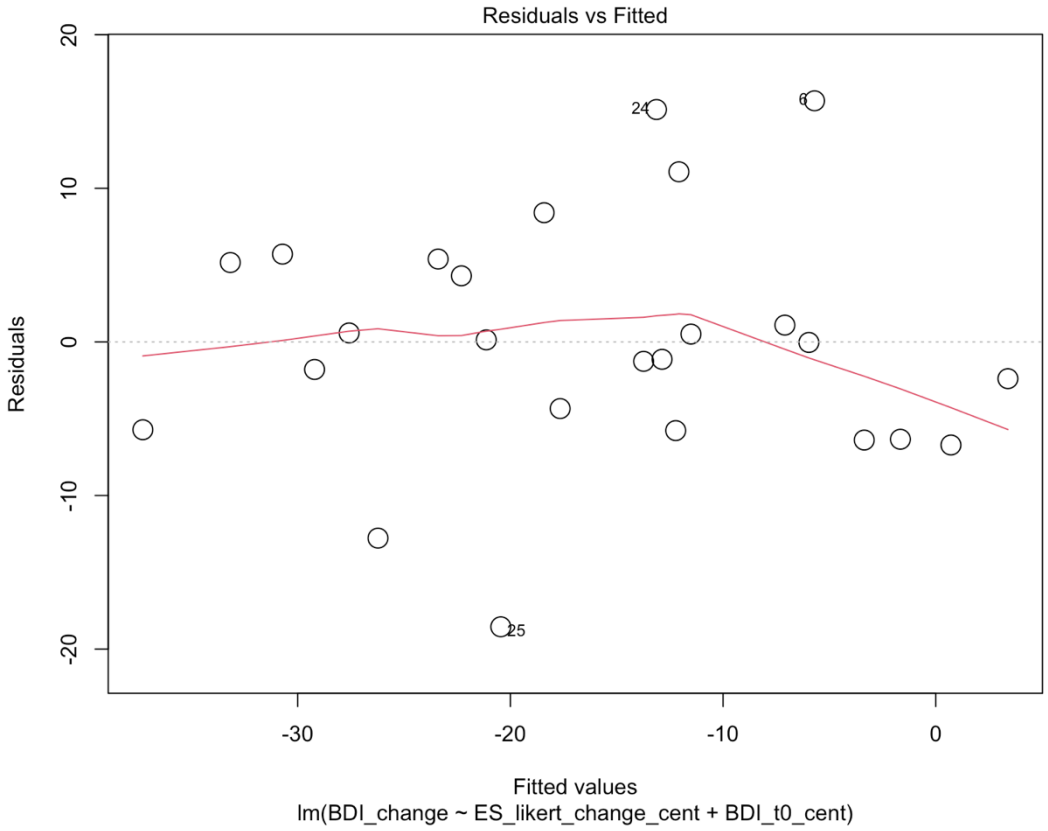


Figure C5

Residuals vs. Fitted Plot for ES-G Model from Incremental Validity Analysis

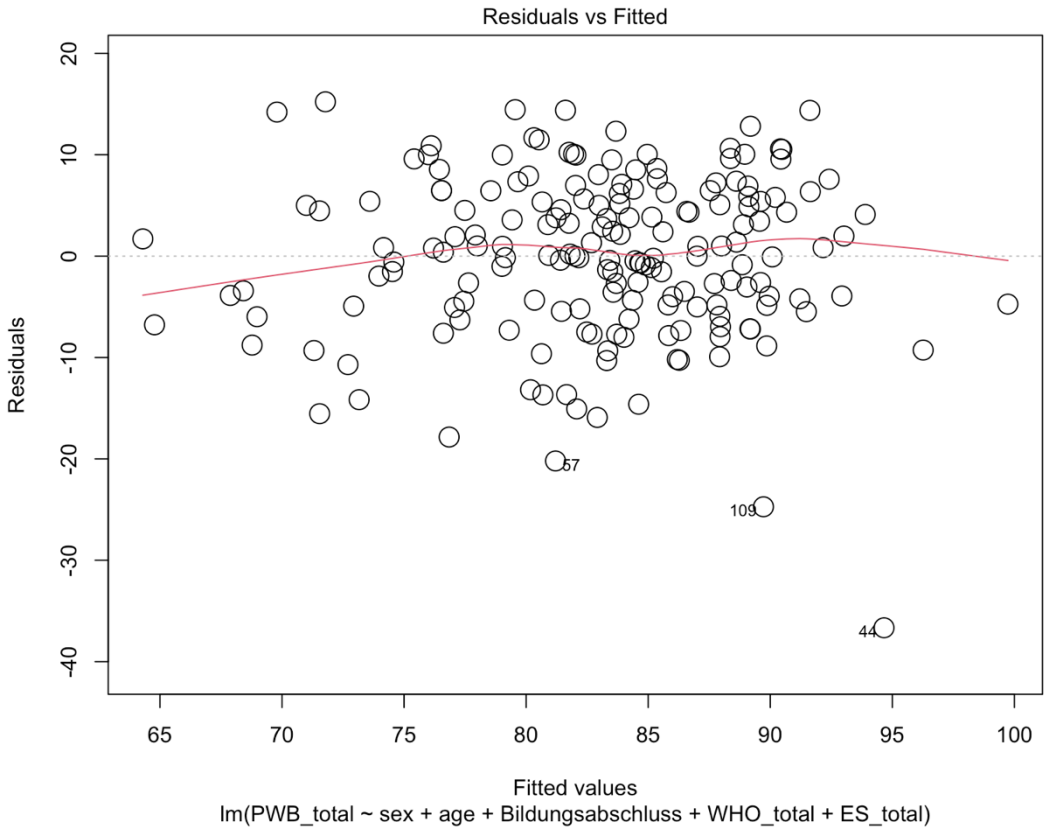
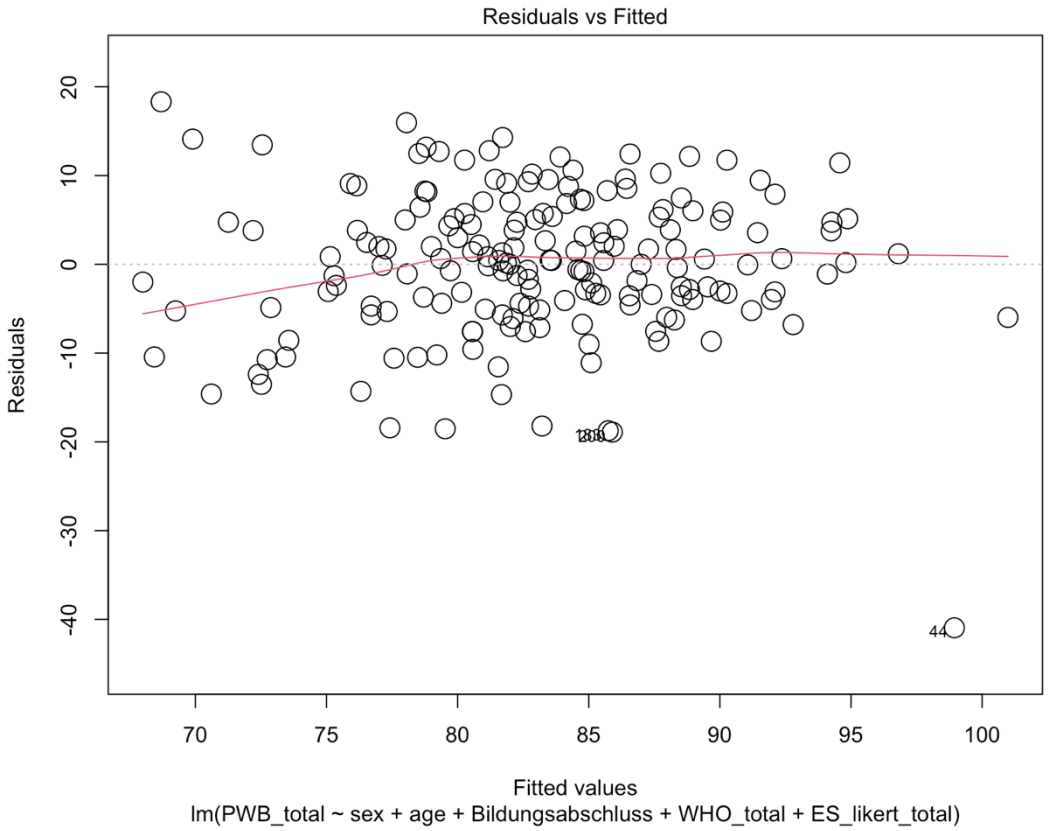


Figure C6

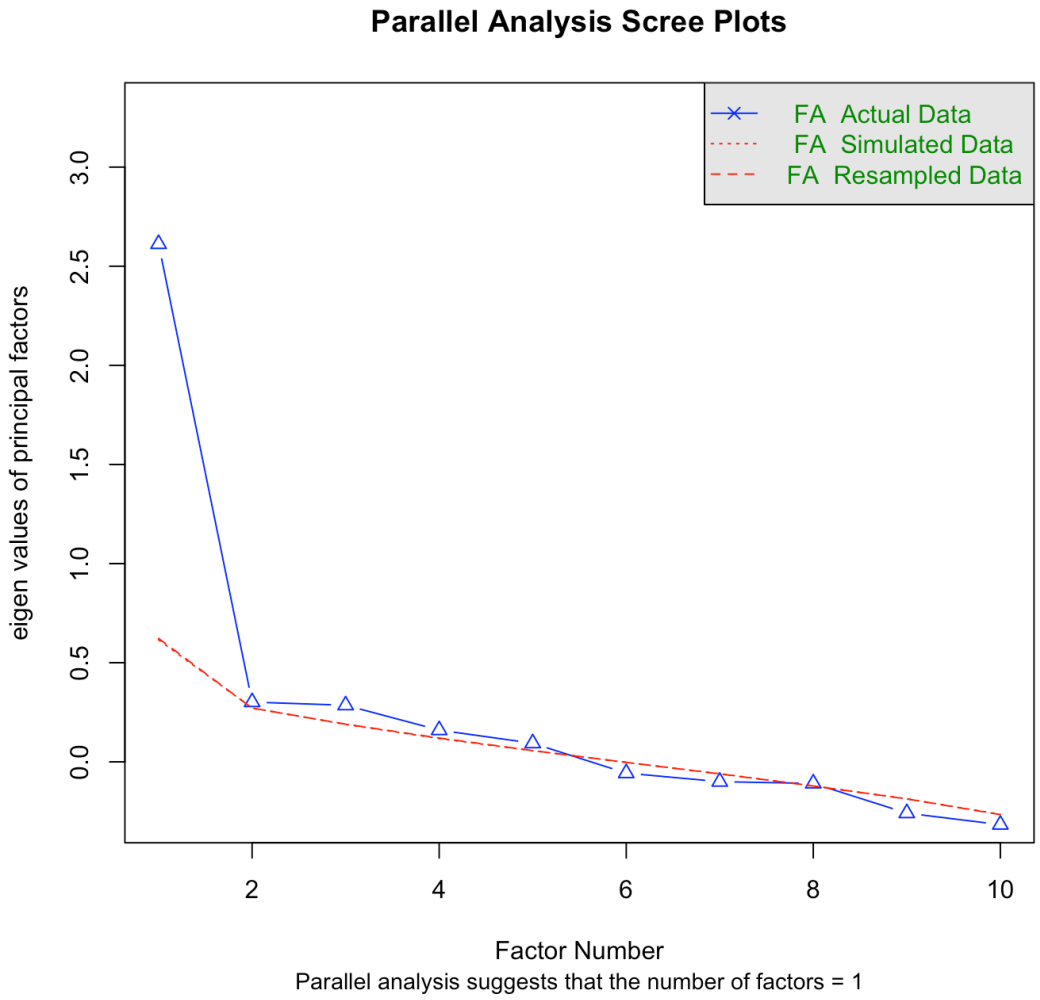
Residuals vs. Fitted Plot for ES-G Likert Model from Incremental Validity Analysis



Appendix D – Parallel Analysis

Figure D1

Parallel Analysis Scree Plot based on Factor Analysis



Note. The blue line represents eigenvalues from the actual data; the red dotted and dashed lines represent the 95th percentile eigenvalues from simulated and resampled data, respectively.

Appendix E – Post-Hoc Comparisons

Table E1

Games-Howell Comparisons of ES-G Total Scores by History of MDE and Current MDE Groups

Group	*n*	Mean (*SD*)	Games-Howell comparison *p*-values			
			0	1	2	3
0	52	8.83 (1.31)				
1	17	7.47 (1.74)	.051			
2	52	7.60 (1.71)	< .001***	.999		
3	44	6.36 (1.66)	< .001***	.189	.005**	
4	41	3.83 (2.25)	< .001***	< .001***	< .001***	< .001***

Note. Group 0 = no history of MDE; Group 1 = full remission; Group 2 = first sub-threshold depressive episode; Group 3 = past MDE + current subthreshold; Group 4 = past MDE + current MDE; *n* = number of participants in each group; Mean (*SD*) = mean values and standard deviations from each group

* $p < .05$; ** $p < .01$; *** $p < .001$

Table E2

Games-Howell Comparisons of ES-G Likert Total Scores by History of MDE and Current MDE Groups

Group	*n*	Mean (*SD*)	Games-Howell comparison *p*-values			
			0	1	2	3
0	52	36.20 (7.53)				
1	17	32.70 (4.09)	.119			
2	52	32.40 (5.97)	< .042*	1.00		
3	44	28.80 (4.93)	< .001***	.026	.012*	
4	41	21.50 (6.87)	< .001***	< .001***	< .001***	< .001***

Note. Group 0 = no history of MDE; Group 1 = full remission; Group 2 = first sub-threshold depressive episode; Group 3 = past MDE + current subthreshold; Group 4 = past MDE + current MDE; *n* = number of participants in each group; Mean (*SD*) = mean values and standard deviations from each group

* $p < .05$; ** $p < .01$; *** $p < .001$

Table E3

Games-Howell Comparisons of ES-G Total Scores by Symptom Severity Groups

Group	*n*	Mean (*SD*)	Games-Howell comparison *p*-values		
			0	1	2
0	132	8.14 (1.46)			
1	28	5.89 (1.40)	< .001***		
2	20	5.20 (2.50)	< .001***	.829	
3	33	3.15 (1.91)	< .001***	< .001***	.013*

Note. Group 0 = minimal depression; Group 1 = mild depression; Group 2 = moderate depression; Group 3 = severe depression; *n* = number of participants in each group; Mean (*SD*) = mean values and standard deviations from each group

* $p < .05$; ** $p < .01$; *** $p < .001$

Table E4

Games-Howell Comparisons of ES-G Likert Total Scores by Symptom Severity Groups

Group	*n*	Mean (*SD*)	Games-Howell comparison *p*-values		
			0	1	2
0	132	34.20 (6.18)			
1	28	28.30 (4.26)	< .001***		
2	20	23.80 (6.86)	< .001***	.151	
3	33	19.50 (6.11)	< .001***	< .001***	.055

Note. Group 0 = minimal depression; Group 1 = mild depression; Group 2 = moderate depression; Group 3 = severe depression; *n* = number of participants in each group; Mean (*SD*) = mean values and standard deviations from each group

* $p < .05$; ** $p < .01$; *** $p < .001$

Appendix F – Receiver Operating Characteristics Curves

Figure F1

ROC Curve for Non-Clinical Group

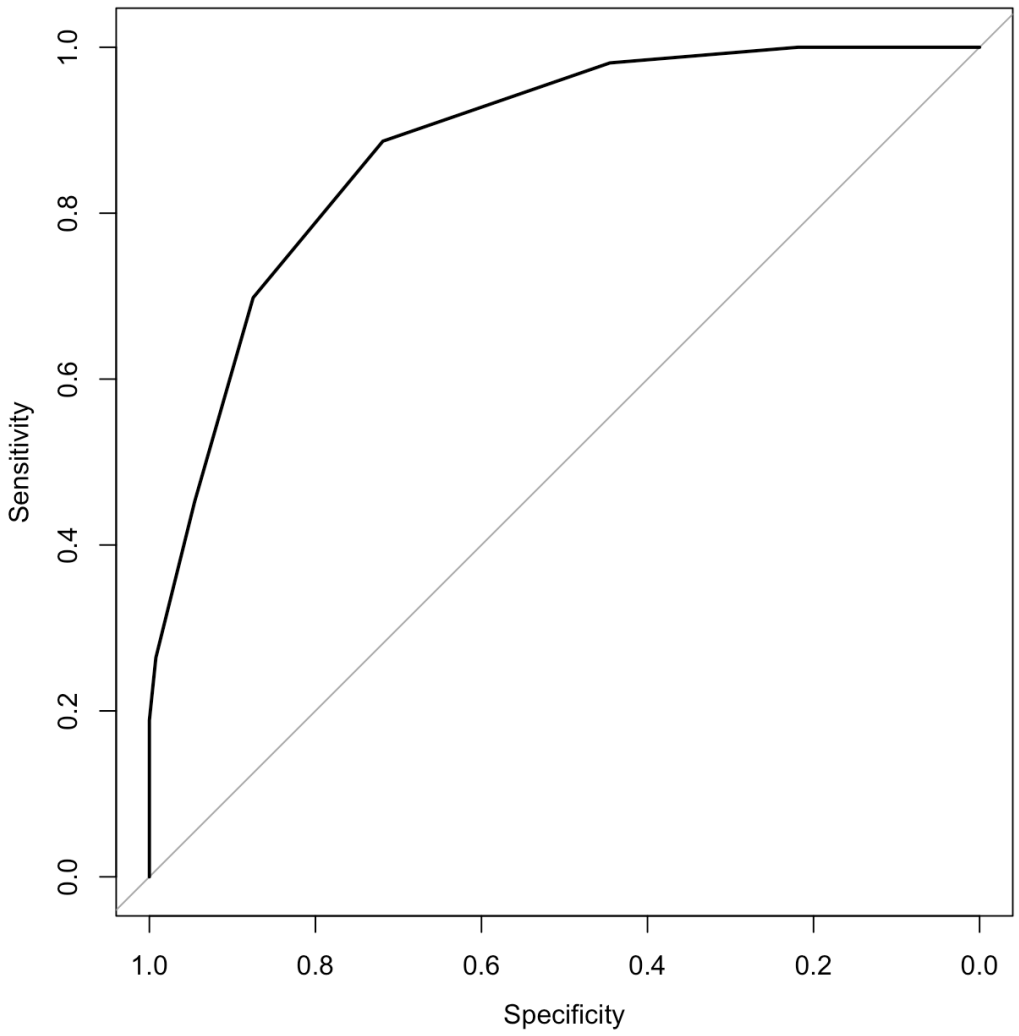
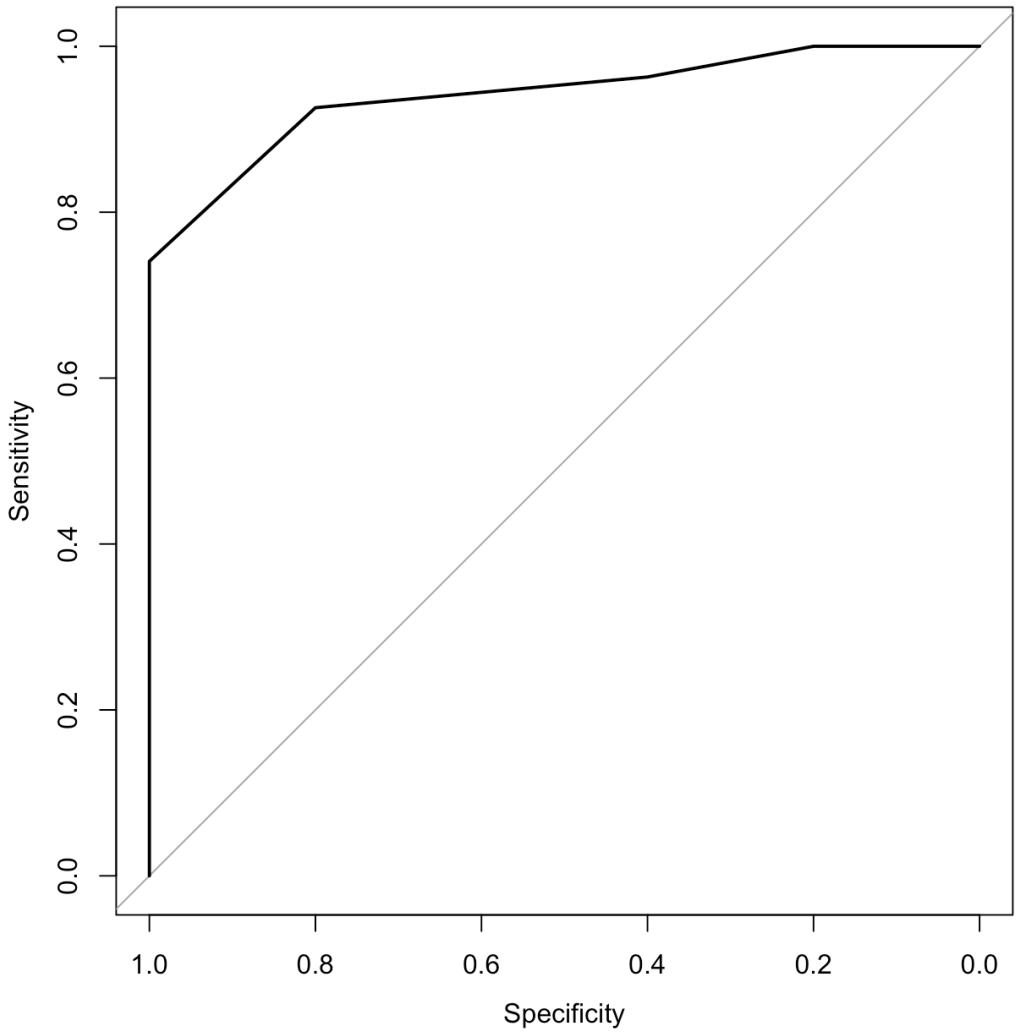


Figure F2

ROC Curve for Clinical Group



Declaration of Authorship

I hereby declare that the thesis submitted is my own unaided work. All direct or indirect sources used are acknowledged as references. Furthermore, this work has not been submitted in the same or a similar form or in part for any other examination.

Munich, XX.XX.XXX