Clinical Psychology Review 119 (2025) 102586

Contents lists available at ScienceDirect

Clinical Psychology Review

journal homepage: www.elsevier.com/locate/clinpsychrev

Evaluation of assessment instruments for working alliance in psychological interventions with adolescents: A systematic review

Mariana Veloso Martins [a], Zorana Jolić Marjanović [b], Nuno Ferreira [c], Camellia Hancheva [d], Emma Motrico [e], Jose M. Mestre [f,g], Nele A.J. De Witte [h], Sibel Halfon [i], Sidse Arnfred [j,k], Margarida Rangel Henriques [a], Nina Petričević [l], Marcin Rzeszutek [m], Jana Volkert [n], Randi Ulberg [o,p], Fredrik Falkenström [q,*]

[a] University of Porto, Faculty of Psychology and Education Sciences (FPCEUP), Rua Alfredo Allen, 4200–135 Porto, Portugal
[b] University of Belgrade, Department of Psychology, Faculty of Philosophy, Čika Ljubina 18-20, 11000 Belgrade, Serbia
[c] University of Nicosia, Department of Social Sciences, School of Humanities and Social Sciences, 46 Makedonitissas Avenue, CY-2417, P.O.Box, 24005, CY-1700, Nicosia, Cyprus
[d] Sofia University "St Kliment Ochridski", 15 Tsar Osvoboditel Blvd, Sofia 1504, Bulgaria
[e] Universidad de Sevilla, Department of Developmental and Educational Psychology, School of Psychology, C/Camilo José Cela, s/n, 41018 Seville, Spain
[f] Universidad de Cádiz, Department of Psychology, Puerto Real, 11519, Spain
[g] Institute University of Social Development and Sustainability (INDESS), Jerez de la Frontera, 11406, Spain
[h] Thomas More University of Applied Sciences, Psychology & Technology, Centre of Expertise Care and Well-being, Molenstraat 8, 2018 Antwerp, Belgium
[i] Istanbul Bilgi University, Psychology Department, santralistanbul, Kazım Karabekir Cad. No: 2/13, 34,060, Eyüpsultan, Istanbul, Turkey
[j] Copenhagen Universityhospital, Region Zealand Mental Health Service, Psychiatric Research Unit, Building 3, Fælledvej 6, 4200 Slagelse, Denmark
[k] University of Copenhagen, Faculty of Health and Medical Sciences, Department of Clinical Medicine, Blegdamsvej 3, 2200 KBH N, Denmark
[l] Teaching Institute of Public Health dr. Andrija Stampar, Department of School and Adolescent health and Youth health Center, Mirogojska 16, 10000 Zagreb, Croatia
[m] University of Warsaw, Faculty of Psychology, Stawki 5/7, 00, –183, Warsaw, Poland
[n] Ulm University, Clinic of Psychosomatic Medicine and Psychotherapy, Albert-Einstein-Allee 23, 89081 Ulm, Germany.
[o] University of Oslo, Institute of Clinical Medicine, Department of Child and adlolescent psychiatry, Postboks 1039 Blindern, 0315 Oslo, Norway
[p] Oslo University Hospital, Mental health and addiction, Child and Adolescent Mental Health Research Unit, Postboks 1039 Blindern, 0315 Oslo, Norway
[q] Linnaeus University, Department of Psychology, Universitetsplatsen 1, SE-352 53 Växjö, Sweden

A B S T R A C T

The working alliance is one of the most robust predictors of outcomes in adult psychotherapy. Since the alliance is often challenging to establish and maintain in psychotherapy with adolescents, conducting high-quality assessments of the alliance using sound measures in this population is critical. Still, measurement instruments developed for adults cannot be directly transferred to adolescent samples. This systematic review aimed to identify and critically evaluate available assessment tools for working alliance in adolescent psychotherapy using the COnsensus-based Standards for the Selection of Health Measurement INstruments (COSMIN) framework. A comprehensive literature search across PsycINFO, PubMed, Web of Science, and PsycARTICLES, up to October 2024, yielded 47 research studies reporting on working alliance measurement properties. Findings indicate that self-report measures are most commonly studied, with the Working Alliance Inventory-Short Form (WAI-S) and Therapeutic Alliance Quality Scale (TAQS) showing the best psychometric properties. Nevertheless, even with these measures, there are notable shortcomings in cross-cultural validity, measurement error, and responsiveness, which are essential for applications in longitudinal studies and with diverse populations. Less commonly studied, often with very small samples, observer-rated tools displayed high reliability but limited predictive validity. Our review highlights the need for more stringent research on developmentally appropriate, reliable working alliance instruments for adolescents to support clinicians and researchers in studying and monitoring this aspect of patient-therapist relations. These findings, together with the COSMIN guidelines, inform recommendations for future research mainly in terms of improved content validity, measurement error, cross-cultural validity, and responsiveness.

1. Introduction

Historically, the concept of working alliance originated in psychoanalysis, where it was defined as the need for the analyst to engage part of the patient's ego as a collaborator in the work of analyzing the patient's internal conflicts (Freud, 1949; Greenson, 1965; Sterba, 1934). Later, Bordin (1979) broadened this concept, making it applicable to any form of psychotherapy. Bordin's definition — still widely accepted

* Corresponding author at: Department of Psychology, Linnaeus University, Universitetsplatsen 1, SE-352 53 Växjö, Sweden.
E-mail address: fredrik.falkenstrom@lnu.se (F. Falkenström).

https://doi.org/10.1016/j.cpr.2025.102586
Received 8 December 2024; Received in revised form 17 March 2025; Accepted 25 April 2025
Available online 27 April 2025

today — suggests that the alliance involves agreement on treatment tasks and goals, supported by a positive emotional bond. The working alliance is psychotherapy's most extensively studied process (Crits-Christoph, Connolly Gibbons, & Mukherjee, 2013), with the latest meta-analysis by Flückiger, Del Re, Wampold, and Horvath (2018) identifying 306 studies examining its relationship with therapeutic outcomes. Although initially controversial (DeRubeis, Brotman, & Gibbons, 2005), today, most researchers across major psychotherapy orientations agree that establishing a working alliance is one of the key conditions of successful psychotherapy. However, some controversies remain regarding which aspects of the alliance are the most critical (e.g., Webb et al., 2011) and whether the alliance directly impacts outcomes or simply facilitates technical interventions (Zilcha-Mano, 2017).

From a developmental psychopathology perspective, adolescence is a critical period for the development of self- (Warschburger et al., 2023) and emotion regulation (Silvers, 2022), yet adolescents often struggle with managing intense emotions due to still-maturing regulatory capacities. Secure relationships, particularly in therapeutic contexts, provide essential scaffolding and co-regulation, helping adolescents navigate emotional challenges until they internalize these skills (Steinberg et al., 2015). The quality of external support from adults and peers significantly influences the development of SR, shaping adolescents' adaptive functioning and resilience (Speranza & Midgley, 2017).

While a secure therapeutic alliance can serve as a vital developmental resource, clinicians have observed that it tends to be fragile (Meeks & Bernet, 2001), requiring careful attention to identify and repair alliance ruptures (Cirasola & Midgley, 2023). Supporting these clinical impressions, research indicates that dropout rates in this age group are high, ranging from 28 to 75 % (de Haan, Boon, de Jong, Hoeve, & Vermeiren, 2013). This may partly be due to the sensitive developmental stage of adolescence, marked by conflicts between the desire for autonomy and a continued need for dependence. Additionally, adolescents' abstract thinking is still developing (Piaget, 1977). These cognitive and emotional factors can make adolescents more prone to temporary breakdowns in their capacity for mentalization (Bleiberg, 2013), which, in turn, heightens the risk of alliance ruptures (Ekeblad, Falkenström, & Holmqvist, 2016). Finally, many adolescents — especially younger ones — do not seek therapy on their own but attend because adults believe it is necessary, adding another layer of complexity to alliance formation in this age group (Koocher, 2003).

Research on the therapeutic alliance in adolescent psychotherapy lags behind that in adult psychotherapy. Karver, De Nadai, Monahan, and Shirk (2018) identified a total of 28 studies examining the alliance–outcome relationship in child and adolescent psychotherapy, a significantly smaller number compared to the 300+ studies available for adult psychotherapy (Flückiger et al., 2018). Furthermore, Karver et al. noted in their review that these 28 studies used 17 different instruments to measure the alliance. This proliferation of measurement tools has also been observed in the adult literature (Flückiger et al., 2018). Elvins and Green (2008) systematically searched for alliance measures and identified no less than 63 instruments developed for adults. Horvath (2018) also observed that the number of alliance instruments tends to grow over time – on the one hand, reflecting an increased awareness of the importance of the therapeutic relationship, but potentially also indicating researchers' discomfort with existing definitions of the alliance as captured by current instruments.

1.1. Measurement issues in relation to working alliance

It could be argued that the therapeutic alliance is, in essence, unobservable, as the quality of collaboration consists of both internal experiences and external interactions. Consequently, various methods for measuring the alliance have been proposed. The most common approach is to ask patients and/or therapists to report their experiences of the alliance through questionnaires. This method has the distinct advantage of being cost-effective and easy to administer, allowing for the collection of large datasets. These datasets, in turn, enable complex statistical analyses of repeated alliance measurements across a broad range of patient-therapist dyads (e.g., Flückiger et al., 2020). However, self-report questionnaires for measuring the alliance come with several limitations. For instance, if each patient completes the alliance measure only once, distinguishing between individual response styles and actual differences in alliance becomes challenging. Factors such as acquiescence bias, social desirability, and personal interpretations of questionnaire items may influence responses. These issues would be considerably less problematic when the alliance is studied by trained observers who evaluate video-recorded sessions. However, while this method addresses several limitations of self-reports, it is considerably more time-consuming and costly. An additional limitation of this approach is that it can only assess externally observable aspects of the alliance, potentially missing the subjective experiences of patient and therapist.

In their review of alliance measures, Elvins and Green (2008) noted that measure development for children and adolescents lags significantly behind that for adults. Given that their review is now 16 years old, an updated review is warranted. Additionally, Elvins and Green did not use a structured system for evaluating the psychometric properties of alliance measures. Since then, the COnsensus-based Standards for the Selection of Health Measurement INstruments (COSMIN; Mokkink et al., 2010) has been established as the most comprehensive frameworks for summarizing and rating psychometric quality. COSMIN was developed by an international team of experts in fields such as epidemiology, psychology, medicine, qualitative research, and health care, with a focus on outcome measurement evaluation.

1.2. The Present Study

In the framework of the European Network on Individualized Psychotherapy Treatment of Young People with Mental Disorders (TREATme; see https://www.cost.eu/actions/CA16102/ and https://www.med.uio.no/klinmed/english/research/networks/treat-me/index.html), we conducted a systematic review of measures of the working alliance in adolescents. We also wanted to rate the quality of these measures using the COSMIN system. The main goal of our review was to analyze and identify the best measures of working alliance for the adolescent population in terms of psychometric properties. We intended to provide recommendations to clinicians and researchers about which measures to use and the gaps that need to be filled by future research.

1.3. Research questions

Our primary research question was: What are the highest quality tools for assessing working alliance in psychological interventions with young people? Specifically, we wanted to 1) identify the existing measures of working alliance in psychological interventions with young people, 2) determine the methodological quality of the studies that report on these measures, and 3) evaluate the quality of the identified measures based on COSMIN criteria.

2. Method

The current review was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement (Moher, Liberati, Tetzlaff, & Altman, 2009). The protocol of the current systematic review was preregistered with the PROSPERO database [CRD42020123317].

2.1. Search strategy

The literature search was conducted using the search engines PsycINFO, PubMed, Web of Science, and PsycARTICLES, covering studies published from inception (dates varied according to the respective

database and go back to 1800) until 28th October 2024. The search was constructed around three categories of search terms: assessment tools, working alliance, and age range (See online supplement Table S14 for a sample search strategy) with the aid of a research librarian. We constructed a search string with working alliance and related synonyms; and age range was specified using search terms and predefined filters in the search engines. The sensitive COSMIN filter search string for measurement properties was used to target assessment tools (Terwee, Jansma, Riphagen, & de Vet, 2009). This filter was constructed to find all studies on the measurement properties of instruments that measure the construct of interest (in this case, working alliance) in the population of interest (youth and young people) in PubMed. The filter has a 97.4 % sensitivity rate and a 4.4 % precision rate compared to hand search. Terwee et al. (2009) also designed an accompanying exclusion filter to remove irrelevant records from the search (e.g., case reports, animal studies). The format of this filter was modified for use in other search engines when necessary. Search citations were reviewed using the Covidence management tool (www.covidence.org), where each reference had to be screened by two independent raters at each stage to evaluate them against the inclusion criteria. Conflicts in the title and abstract screening were handled with an over-inclusive approach, obtaining full-text articles for further investigation. M.M., N.D.W., F.F., and Z.J.M. solved final conflicts regarding inclusion. Finally, reference lists of the included articles were hand-searched by J.M.M., E.M., S.H., N.P. and C.H. for additional relevant studies.

2.2. Inclusion criteria

To be included, an article had to: (1) be published in a peer-reviewed journal; (2) be written in English; (3) include a measure of therapeutic/working alliance between a young patient and a mental health professional; (4) have that measure applied to adolescents (patient/therapist-reported or observer-rating), age range between 12 and 19 years or with a sample mean age between 12 and 19. The age range of 12 to 19 years was chosen to ensure the inclusion of studies focusing on adolescents, aligning with the American Psychological Association definition of adolescence as beginning with puberty and ending with physiological maturity around age 19 (VandenBos, 2015). Studies using measures that assessed working alliance through related constructs (e.g. empathy) or as a subscale from a different measure were excluded. However, previously translated, shortened, or altered measures were included and analyzed. Besides analyzing the original scales to support data extraction, further information was sought to support data extraction from the authors (via email) or previously referenced manuscripts or reports. Two members of the research team independently made decisions about the inclusion/exclusion of studies, and any disagreements were resolved through consultation with a third reviewer.

2.3. Data extraction

Data from the selected articles were extracted to assess study quality and the circumstances in which the instrument was used. The following information was collected from the articles: sample size, age of the sample, clinical status of the sample, setting in which it was applied, number of therapists, timing of assessment(s), number of items and response options used, and psychometric properties studied, and language of the instrument. This information can be found in the online supplement (Tables S1-S2).

2.4. Assessment of measurement properties of included studies

Measurement properties were critically appraised using the COSMIN taxonomy (Mokkink et al., 2010). For the present study, we used: a) the COSMIN criteria for evaluating the content validity of health-related Patient Reported Outcomes (Terwee et al., 2018); b) the COSMIN guideline for systematic reviews of measurement instruments (Prinsen et al., 2018), including the supplement for systematic reviews of Patient-Reported Outcome Measures; c) the COSMIN Risk of Bias checklist (Mokkink et al., 2018) including the modified Grading of Recommendations Assessment, Development and Evaluation (GRADE) for grading the quality of evidence. COSMIN manuals can be consulted at www.cosmin.nl/index.html. Measurement properties are organized according to three domains: i) reliability, i.e. the extent to which the instrument is free from measurement error (including internal consistency, reliability, and measurement error); ii) validity, i.e. the extent to which the instrument measures the construct originally proposed (including content validity, construct validity, and criterion validity); and iii) responsiveness, i.e. the extent to which the instrument assesses changes in the underlying construct in a valid way. We decided not to include the measurement property criterion validity in this systematic review since this requires comparison to a gold standard instrument, and no such gold standard exists for adolescent working alliance. This taxonomy is transferred into 5 to 18 items for each measurement property assessing aspects of study design and statistical analyses. These risk of bias items can be scored on a 4-point rating scale (very good, adequate, doubtful, and inadequate), which allows for calculating a methodological quality score based on the lowest rating for each item (Terwee et al., 2012). All team members received introductory COSMIN training, and then each instrument was independently reviewed by two research team members. For patient-reported measures, MM reviewed all studies together with at least one other team member. FF reviewed all studies using observer-rated measures together with at least one other team member. In a final round, FF reviewed all ratings (self- and observer-rated) to check that assessments were similar across all studies.

3. Results

Fig. 1 shows a Prisma flowchart of the inclusion/exclusion process. In total, 47 papers were included, with 36 reporting on self-report measures and 13 on observer measures (two studies reported on both self-report and observer measures). All studies except one (from China) were conducted in high-resource settings. Tables 1 and 2 summarizes the evidence for psychometric properties of self-report and observer measures, respectively. Tables with information on all included studies are available in the online supplement Tables S1-S2, and tables with detailed information on psychometric properties is available in online supplement Tables S3-S13.

Self-report measures were the most common; our review identified 15 self-report measures studied across 36 studies. In comparison, there were seven instruments and 13 studies of observer measures. As expected, the sample sizes were also considerably larger for self-report measures than for observer measures ($M_{self\text{-}report} = 177$, $M_{observer} = 46$). The psychometric properties most frequently studied were internal consistency for self-report measures, analyzed in 29 of the 36 studies, and inter-rater reliability for observer measures, analyzed in 12 of the 13 studies. However, some properties identified as important by COSMIN—such as content validity, measurement error, and cross-cultural validity/measurement invariance—were almost entirely lacking.

3.1. Self-report Measures

3.1.1. Description of Included Studies

Characteristics of self-reported included studies are shown in Table S1. The review identified 15 instruments used in 36 studies. In total, the 36 studies assessed 6012 adolescents' self-reported alliance. Most of the participants were female (66 %), with 30 studies reporting on gender. With the exception of one study from Australia (Anderson et al., 2012), one from Israel (Mekori-Domachevsky et al., 2023) and one from China (Hou et al., 2024) the studies were conducted in Europe and North America. The country with the most studies ($n = 12$) was the USA (see Table S1 in the online supplement). For eight of the 15 measures, there was also information on the respective therapist versions

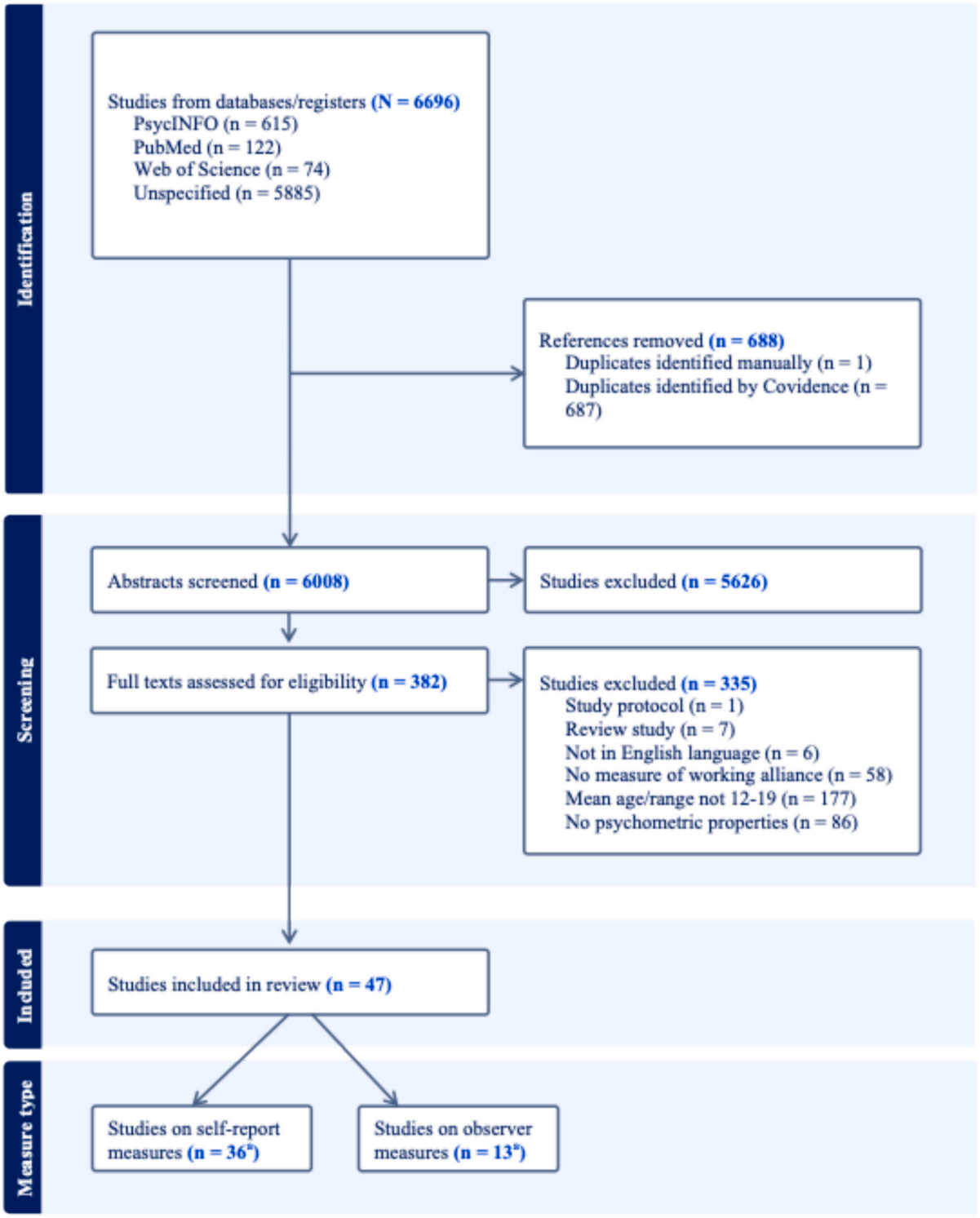

Fig. 1. PRISMA flow diagram.

(Table S2). These were reported in 12 studies that depicted results on both adolescents' and therapists' reports of working alliance. The number of therapists ranged between two and 713.

Following COSMIN criteria, modified versions of measures were reviewed separately from each other. This was the case with the original Therapeutic Alliance Scale for Children (TASC; Shirk & Saiz, 1992), which was shortened from 13 to 12 items into Therapeutic Alliance Scale for Children–revised (TASC-r; Creed & Kendall, 2005); and with the Working Alliance Inventory (Horvath & Greenberg, 1989), which was shortened from 36 to two different versions with 12 items each: the Working Alliance Inventory – Short form (WAI—S; Tracey & Kokotovic, 1989), and the Working Alliance Inventory – Short form Revised (WAI-SR; Hatcher & Gillaspy, 2006). Some studies had overlapping samples but assessed different psychometric properties; with three studies from the Impact study using the WAI-S (Cirasola et al., 2022; Cirasola, Midgley, Fonagy, Impact Consortium, et al., 2021; Cirasola, Midgley, Fonagy, & Martin, 2021) and two studies on the child version of the Session Rating Scale (Hauber & Boon, 2022; Hauber, Boon, &

Table 1
Summary table of measurement properties of self-report measures of working alliance in adolescents.

Measure	Structural validity		Internal consistency		Reliability		Concurrent vaidity		Predictive validity		Responsiveness	
	Pooled rating	Quality	Pooled rating	Quality	Pooled rating	Quality	Pooled rating	Quality	Pooled rating	Quality	Pooled rating	Quality
C-SRS			+	Low					−	Moderate	+	Moderate
FTAS	−	Very low										
I-TAS	+	Moderate	+	Moderate			−	Low				
SOFTA-s patient			+	Moderate	+	Very low					−	Low
SOFTA-s therapist			+	Low	+	Very low			−	High	+	Low
SOFTA-s(i)			+	Low							−	Very low
TAQS	+	High	+	High					−	High	−	High
TASC			+	Very low	+	Very low			−	Low	−	Low
TASC-r	+	Low	+	Moderate			−	Moderate	−	High		
WAI-S patient	+	High	+	High	−	Low			+	High		
WAI-S therapist	+	High	+	High	−	Low			−	High		
WAI-SR patient			+	Moderate					−	Moderate		
WAI-SR therapist			+	Low					−	Low		
SAI patient			+	Low							+	High
SAI therapist			+	Low							+	High
FTASr-SF			+	Low	−	Very low	+	Low	−	High		
IAM-F	+	Moderate	+	High	+	Very low	+	Low	+	High		
HAQ	+	Moderate	+	High	+	Moderate			+	Low		

Note. C-SRS = Child version of the Session Rating Scale; FTAS = Family Therapy Alliance Scale; I-TAS = Inpatient-Treatment Alliance Scale; SOFTA-s = System for Observing Family Therapy Alliances-self report; SOFTA-s(i) = System for Observing Family Therapy Alliances, individual therapy version; TAQS = Therapeutic Alliance Quality Scale; TASC = Therapeutic Alliance Scale for Children; TASC-r = Therapeutic Alliance Scale for Children –revised; WAI = Working Alliance Inventory; WAI-S = Working Alliance Inventory – Short form; WAI-SR = Working Alliance Inventory – Short form Revised; SAI = Session Alliance Inventory; FTASr-SF = Family Therapy Alliance Scale revised short form; IAM-F = Intersession Alliance Measure - Family version; HAQ = Helping Alliance Questionnaire.

Table 2
Summary table of measurement properties of observer measures of working alliance in adolescents.

Measure	Structural validity		Internal consistency		Reliability		Concurrent validity		Predictive validity	
	Pooled rating	Quality	Pooled rating	Quality	Pooled rating	Quality	Pooled rating	Quality	Pooled rating	Quality
VTAS-R	+	Low	+	High	+	High			−	High
VTAS-R-SF	+	Low	+	High	+	High	+	Low	−	High
ATAS	+	Very low	+	Very low	+	Moderate	+	Moderate	−	Moderate
WAI-O			+	Very low	+	Moderate	+	Low	−	Low
TPOCS-A	+	Low	+	High	+	High	+	High	−	Moderate
SOFTA-O			+	Very low					−	Very low
AOCS			+	Very low	+	Very low	+	Low	+	Low

Note. VTAS-R = Vanderbilt Therapeutic Alliance Scale - Revised; VTAS-R-SF = Vanderbilt Therapeutic Alliance Scale – Revised Short Form; ATAS = Adolescent Therapeutic Alliance Scale; WAI-O = Working Alliance Inventory – Observer version; TPOCS-A = Therapeutic Process Observational Coding System for Child Psychotherapy – Alliance Scale; SOFTA-O = System for Observing Family Therapy Alliance – Observer version.

Vermeiren, 2020). One study reported on two measurement instruments (Anderson et al., 2024).

The WAI-S (Tracey & Kokotovic, 1989) was the most used self-report measure, with nine studies using it to assess adolescent-reported alliance (Cirasola et al., 2022; Cirasola, Midgley, Fonagy, Impact Consortium, et al., 2021; Cirasola, Midgley, Fonagy, & Martin, 2021; Hawley & Garland, 2008; McLeod et al., 2024; Rienecke, Richmond, & Lebow, 2016; Rollin, Pascuzzo, & Lanctot, 2024; van Benthem et al., 2020; van Benthem et al., 2024). This was followed by the WAI-SR (Hatcher & Gillaspy, 2006), which was used in six studies (Cooper, Connor, Orloff, Herrington, & Timko, 2024; Diamond et al., 2024; Hou et al., 2024; King, Richner, Tuliao, Kennedy, & McChargue, 2019; Mekori-Domachevsky et al., 2023; Nissling et al., 2023). The original 36-item WAI was used in three studies (Ayotte, Lanctot, & Tourigny, 2016, 2017; Gergov, Marttunen, Lindberg, Lipsanen, & Lahti, 2021), although all these studies modified the scale by reducing the number of items. The TASC-r (Shirk & Saiz, 1992) was used in four studies (Jacoby et al., 2021; Ormhaug, Shirk, & Wentzel-Larsen, 2015; Ovenstad, Ormhaug, & Jensen, 2023; Ovenstad, Ormhaug, Shirk, & Jensen, 2020), the System for Observing Family Therapy Alliances-self report (SOFTA-s; Friedlander, Escudero, & Heatherington, 2006) was used in two studies (Jewell et al., 2023; Kivlighan Jr., Escudero, Friedlander, & Orlowski, 2022), and the Helping Alliance Questionnaire (HAQ; Alexander & Luborsky, 1986) was studied by Kermarrec, Kabuth, Bursztejn, and Guillemin (2006) and Steil, Weiss, Renneberg, Gutermann, and Rosner (2023). We found one study of the child version of the Session Rating Scale (C-SRS; Duncan et al., 2003), which was reported on in two publications (Hauber et al., 2020; Hauber & Boon, 2022).

We further identified eight measures that were used in one study each: Johnson, Ketring, and Anderson (2013) applied the Family Therapy Alliance Scale (FTAS; Pinsof & Catherall, 1986); Anderson et al. (2024) studied both the Family Therapy Alliance Scale revised short form (FTASr-SF; Pinsof, Zinbarg, & Knobloch-Fedders, 2008) and the Intersession Alliance Measure, family version (IAM—F). Haggerty et al. (2015) assessed alliance with the Inpatient-Treatment Alliance Scale (I-TAS; Blais, 2004); Bickman et al. (2012) applied the Therapeutic Alliance Quality Scale (TAQS; Bickman et al., 2010); Kang et al. (2021) used the Therapeutic Alliance Scale for Children (TASC; Shirk & Saiz, 1992);

and Alvarez, Herrero, Martínez-Pampliega, and Escudero (2021) examined the individual therapy version of the SOFTA-s. Finally, Lindqvist et al. (2023) studied the Session Alliance Inventory (SAI; Falkenström, Hatcher, Skjulsvik, Larsson, & Holmqvist, 2015). With the exception of the TAQS (Bickman et al., 2010), all measures were initially developed for an adult population.

The working alliance was assessed at different time-points, ranging from the first (Hauber et al., 2020; Hawley & Garland, 2008; van Benthem et al., 2020) to the last session (Haggerty et al., 2015; Johnson et al., 2013), or in one study even at follow-up (Kermarrec et al., 2006). The number of repeated measures of alliance ranged from one to ten.

3.1.2. Content Validity

Four studies reported assessing at least one aspect of content analysis, in four different measures (see detailed summary and ratings in Suppl. Table 1). While three of these studies (Anderson et al., 2012; Haggerty et al., 2015; Kermarrec et al., 2006) assessed measures developed for the adult population (HAQ, I-TAS and WAI—S), only Bickman et al. (2012) used a scale specifically developed for the youth population (TAQS).

We found no self-report measure that consulted adolescents on the items' relevance, comprehensiveness, or comprehensibility regarding the measures of working alliance, one of the COSMIN quality indices for content validity. This includes the original studies focusing on the development of these measures. For example, Horvath and Greenberg (1989) performed content analyses of the WAI with young adults (postgraduates), but not adolescents. Two studies (Anderson et al., 2012; Haggerty et al., 2015) refer to some degree of rewording to accommodate the adolescent population, but these were rated as inadequate due to insufficient evidence (no reference to the method used or professionals consulted, which is needed in the COSMIN framework). Both the French version of the HAQ (Kermarrec et al., 2006) and the TAQS (Bickman et al., 2012) were rated as adequate because the development of the measures used appropriate criteria in gathering information from professionals. However, overall, content validity was rated as indeterminate, with moderate quality of evidence due to the lack of input from the target population. We conclude that the HAQ and TAQs showed the best evidence for content validity. However, the evidence is based on only one study each, and there was no participatory research involving adolescents.

3.1.3. Structural Validity

Structural validity was studied for eight measures, in 11 studies. The WAI-S was the most studied instrument (three studies; Anderson et al., 2012; Cirasola, Midgley, Fonagy, & Martin, 2021; van Benthem et al., 2024), with methodological quality rated from 'adequate' to 'very good'. None of these studies supported the theoretical three-factor structure (agreement on goals, tasks, and emotional bond). According to COSMIN criteria, the one-factor model was supported in all three studies; and taken together, the three studies provide high confidence in the findings. The TAQS (Bickman et al., 2010) was analyzed in one relatively large study ($N = 679$; Bickman et al., 2012) which was rated as with very good methodological quality, resulting in high confidence in the one-factor structure. The factor structures of FTAS, HAQ, I-TAS, TAQS, TASC-r, IAM—F, and the original WAI were all tested in one study each, with methodological quality varying from very low to moderate.

3.1.4. Internal Consistency

Internal consistency was the most studied psychometric property. Twenty-eight studies reported internal consistency for 14 measures. Internal consistency was almost always high; however, the COSMIN system emphasizes that internal consistency for a sum score requires at least low support for unidimensionality. Otherwise, internal consistency needs to be reported separately for each subscale. Failing to do so, or deficient support for unidimensionality from structural validity research were the most common reasons for downgrading the methodological quality of a study. Again, the WAI-S was studied the most, with five studies reporting on its internal consistency (Anderson et al., 2012; Cirasola, Midgley, Fonagy, & Martin, 2021; Hawley & Garland, 2008; McLeod et al., 2024; van Benthem et al., 2024). These studies all showed good to excellent internal consistency for the total score. Since the WAI-S had shown unidimensionality (see previous section), they were all rated as very good in terms of methodological quality. Therefore, the confidence in the internal consistency of the WAI-S was rated as high.

The internal consistency of the WAI-SR was evaluated in five studies (Diamond et al., 2024; Hou et al., 2024; King et al., 2019; Mekori-Domachevsky et al., 2023). Since the structural validity of the WAI-SR has not been studied in this population, these studies were all rated as of doubtful methodological quality, and the quality of the evidence for internal consistency of the WAI-SR was therefore rated as moderate despite all studies showing high internal consistency for total scores and, in one study, subscales. The original WAI was explored in four studies (Ayotte et al., 2016, 2017; Rollin et al., 2024), although all of these studies modified the scale by eliminating some of the items before calculating internal consistency, which makes any pooled conclusion misleading. All these studies calculated internal consistency for the total score, and due to absence of structural validity evidence of unidimensionality, these were therefore rated as of doubtful quality.

The TASC-r was assessed in four studies (Jacoby et al., 2021; Ormhaug et al., 2015; Ovenstad et al., 2023; Ovenstad, Jensen, & Ormhaug, 2022). Again, these studies all calculated internal consistency for the total score. The only study reporting on structural validity for the TASC-r (Ormhaug et al., 2015) found a two-factor structure while not reporting the fit of the one-factor model. Therefore, these studies were all rated as of doubtful methodological quality, and the pooled rating, although positive, was deemed only moderate evidence for the internal consistency of the TASC-r.

Two articles (Jewell et al., 2023; Kivlighan Jr. et al., 2022) reported the internal consistency for the SOFTA-s (Friedlander et al., 2006). Again, since the factor structure of this instrument has not been evaluated for this population, both these studies were rated as of doubtful methodological quality, and the quality of evidence as moderate for patient ratings and low for therapist ratings (since therapist ratings were only included in one of the studies). The HAQ, TAQS, and IAM-F were all explored in one study each, and all were rated as of very good methodological quality and therefore the positive internal consistency was rated as high quality of evidence. The remaining studies were downgraded either for small sample size (Haggerty et al., 2015) or because structural validity evidence was not available for this population. The latter was the case for the FTASr-SF (Anderson et al., 2024), SOFTA-s(i) (Escudero, Friedlander, Kivlighan, Orlowski, & Abascal, 2022), C-SRS (Hauber et al., 2020), and SAI (Lindqvist et al., 2023). Finally, the study of TASC (Kang et al., 2021) was considered very low evidence due to both small sample size and unclear structural validity in this population.

3.1.5. Reliability

Seven articles reported information on the stability of alliance measures over time. This information was sometimes reported as test-retest reliability, while in some articles, we extracted this information despite the authors not having intended it as test-retest reliability. For test-retest reliability, COSMIN requires 1) that patients are stable in the interim between measures, and 2) that the time interval is appropriate. We decided that for working alliance to be rated as 'very good' on these aspects, the time interval should be between sessions or one week. While the criterium of the next session was chosen because we cannot assume the alliance is stable as soon as another session has occurred between measures. The option of having also one week was important because some interventions involve multiple components beyond just individual or group therapy sessions, meaning that the appropriate time interval may be determined by the structure of the intervention rather than a fixed number of days. Moreover, COSMIN requires that intraclass correlation rather than Pearson *r* or Spearman *rho* be calculated (or

weighted Kappa for categorical ratings) for a 'very good' rating.

Methodological quality was deemed inadequate in all studies except one (Kermarrec et al., 2006), which was rated as adequate. This was due to several factors: patients were not stable, the time interval was inappropriate, and Pearson correlations (rather than ICC) were reported. Therefore, the quality of evidence was rated from very low to moderate, with the evidence for HAQ rated as moderate.

3.1.6. Measurement Error

In repeated measures designs, it is possible to calculate the standard error of measurement (SEM), which, according to COSMIN, is the preferred statistic for measurement error. In psychotherapy research, reporting the reliable change index (RCI; Jacobson & Truax, 1991) is relatively common. The RCI builds on the SEM, and is a way of quantifying the smallest detectable change (over and above measurement error). In the COSMIN system, calculating the SEM based on internal consistency is considered inadequate since this does not take the variance over time into account. This means that the two studies included in our review that calculated the SEM (Bickman et al., 2012; Hauber et al., 2020) were rated as of inadequate methodological quality. Moreover, for evaluating the RCI in a psychometric study, the COSMIN standard is that the minimally important change has been defined, and this was not done in any of the studies in our review. Therefore, the results regarding measurement error were considered indeterminate, and very low quality of evidence.

3.1.7. Construct Validity and Responsiveness

Construct validity is a broad term encompassing all other aspects of validity (Borsboom, Mellenbergh, & van Heerden, 2004; Cronbach & Meehl, 1955). In this review, we have divided construct validity into comparisons with other alliance measures (concurrent validity) and predictions of outcome (predictive validity). We followed the COSMIN criterion that requires that hypotheses are tested, and a '+' rating is given only when at least 75 % of hypotheses regarding construct validity are supported. However, we could not follow the COSMIN recommendation to use effect sizes rather than statistical significance since many studies did not report standardized effect sizes. Therefore, we rated a hypothesis as having been supported if the test showed a statistically significant effect.

Using these criteria, only two measures, the FTASr-SF and IAM—F, were rated as having shown concurrent validity, and the evidence for both was rated as low. Likewise, for predictive validity, most measures did not reach the criterion of having at least 75 % of hypotheses supported. Only the WAI-S and IAM-F were rated as showing positive predictive validity with high quality evidence. The HAQ also showed positive predictive validity, although with low quality of evidence due to the very small sample size of the supporting study ($N = 38$; Steil et al., 2023).

COSMIN defines responsiveness as the construct validity of change scores. In this review, we classified studies on within-patient time-lagged predictions as providing evidence for responsiveness due to the fact that within-patient effects can be shown to be mathematically equivalent to change scores (Usami, Murayama, & Hamaker, 2019). Given this, of the included measures, only the SAI showed positive results for responsiveness with high quality evidence (Lindqvist et al., 2023). The C-SRS and SOFTA-s therapist form also showed positive results for responsiveness, with moderate and low quality of evidence, respectively. None of the other five measures that provided any tests of responsiveness reached the criterion of having at least 75 % of hypotheses confirmed and were thus rated as negative.

3.2. Observer Measures

3.2.1. Description of Included Studies

Characteristics of the included studies of observer measures are shown in Table S2 in the online supplement. Our review identified seven instruments used in 13 studies. In total, the 13 studies assessed alliance in 646 adolescents' therapies. Most of the participants were male (56 %). Of the 13 included studies, nine were from the USA, two from Norway, one from Germany and one from the UK. None of these studies included information on cross-cultural validity, measurement error, or responsiveness.

3.2.2. Content Validity

Content validity was not formally studied in any of the included studies on observer measures. Since content validity for observer-rated alliance measures does not quite fit the COSMIN system, which is intended for patient-rated outcome measures, we below briefly describe the information we have on the development of the included measurement instruments:

The ***Vanderbilt Therapeutic Alliance Scale (VTAS;*** Hartley & Strupp, 1983***)*** is a 44-item, observer-rated instrument designed to measure the strength of the therapeutic alliance in adult individual therapy. The revised VTAS (VTAS-R) includes 24 items taken from the client and therapist-client interaction scales — the therapist contribution scale was eliminated because of its overlap with therapist techniques — and has some items slightly reworded for a better fit with treatment involving adolescents and families. The five-item short form (VTAS-R-SF), developed by Shelef and Diamond (2008), included some minor revisions to clarify items or improve interrater reliability.

The ***Adolescent Therapeutic Alliance Scale (***Johnson, Hogue, Diamond, Leckrone, & Liddle, 1998***)*** is an observer-rated tool designed to assess key dimensions of the therapist-adolescent working alliance across diverse counseling contexts. In selecting factors for inclusion, the theoretical and empirical literature on alliance scales used with both adolescent and adult samples was consulted. The 44-item VTAS provided a foundation for the scale's development. Items were revised or removed to ensure applicability across the adolescent age range, adaptability to various intervention settings, and sensitivity to developmental issues relevant to adolescents. Some items were eliminated or rephrased for developmental appropriateness (Faw, Hogue, Johnson, Diamond, & Liddle, 2005).

The ***Working Alliance Inventory – Observer version (WAI—O;*** Darchuk et al., 2000***)*** is the same scale used in the adult working alliance literature. It was used in one study included in our review (Puls, Schmidt, & Hilbert, 2019), with no information on any adaptations to the adolescent population.

The ***Therapy Process Observational Coding System-Alliance Scale (TPOCS-A;*** McLeod & Weisz, 2005***)*** is a nine-item measure consisting of a child and parent form designed to objectively describe the child–therapist and parent–therapist alliance. A literature review was used to identify relevant dimensions (bond and tasks), followed by a sampling of items from existing alliance scales. These items were then pilot-tested and refined to improve inter-rater reliability.

System for Observing Family Therapy Alliances - Observational measure (SOFTA-O; Friedlander et al., 2006***)***. The SOFTA consists of four subscales: Engagement in the Therapeutic Process, Emotional Connection to the Therapist, Safety within the Therapeutic System, and Shared Sense of Purpose within the Family. To develop the SOFTA, the researchers reviewed existing alliance measures and conducted a comprehensive search of theoretical and empirical literature to identify specific descriptors of the therapeutic relationship in couple and family therapy (Friedlander et al., 2006). Drawing on this literature and their own clinical experiences, they generated a pool of both positive and negative items representing various aspects and levels of client collaboration in family therapy. The researchers then observed 12 videotaped family sessions, for which clients had provided self-reported perceptions of the alliance. They identified individual and interpersonal behaviors aligning with clients' reported positive or negative experiences in sessions. The authors refined the scale by evaluating the items' face validity: they asked family therapy process researchers from the United States, Canada, and Spain to review the dimension definitions and select

the construct best represented by each item. Items were retained if at least 75 % of experts agreed on their dimensional relevance. Finally, the scale was tested for inter-rater reliability and known-groups validity using vignettes of sessions with high versus low alliance.

The ***Alliance Observation Coding System (AOCS;*** Karver, Shirk, Day, Field, & Handelsman, 2003***)*** was designed to consider adolescent development and adolescent emotional and cognitive responses to a therapist in individual therapy. The system was designed after the authors had reviewed the attachment and social development literature, interviewed clinicians who worked with adolescents, and drew from their experiences working extensively with adolescents in individual therapy (Karver et al., 2008).

In summary, content validation methods varied across instruments. The WAI-O seems to have been adopted from adult psychotherapy without adaptation, while the VTAS-R and VTAS-R-SF incorporated some minor modifications. The ATAS appears to have undergone more extensive adaptation. In contrast, the TPOCS-A, SOFTA, and AOCS were specifically developed considering adolescent populations. Scale development was primarily based on expert opinions, with the SOFTA being a partial exception. None of the scales reported involved direct interviews with clinicians or patients, as recommended by the COSMIN criteria.

3.2.3. Structural Validity

Five papers used exploratory factor analysis (principal component or principal axis factoring) to study the structural validity of four measures: the VTAS-R (Robins et al., 2003; Hogue, Dauber, Stambaugh, Cecero, & Liddle, 2006), VTAS-R-SF (Shelef & Diamond, 2008), ATAS (Faw et al., 2005), and TPOCS-R (Fjermestad et al., 2012). All studies except one (Hogue et al., 2006) had sample sizes deemed too small for factor analysis (N ⟨100). All measures showed single-factor solution, except for one study, which was rated as inadequate methodological quality due to the low sample size in relation to the number of items on the scale. The evidence for structural validity was deemed low (VTAS-R, VTAS-R-SF, and TPOCS-R) to very low (ATAS).

3.2.4. Internal Consistency

Internal consistency was reported in eleven papers studying six measures (VTAS-R, VTAS-R-SF, TPOCS-A, AOCS, WAI—O, and ATAS). The reported alphas were all in the good to excellent range (> 0.80). For three measures, VTAS-R, VTAS-R-SF, and TPOCS-A, the methodological quality of the supporting studies was deemed very good, and the quality of evidence for the measures was high. The AOCS, WAI—O, and ATAS were all rated as having moderate quality of evidence due to the uncertainty regarding structural validity and the small sample sizes of the studies.

3.2.5. Reliability

Twelve studies on six measures reported reliability; all but one used intra-class correlation (the other study used weighted Kappa for ordinal data). Eight studies reported reliability for the total scale, while four reported item-level reliability. The reported ICCs were all acceptable to excellent (>0.70 for the total scale). Individual study quality for reliability was rated very good for eight studies of twelve. The most common reason for lower scores was not reporting the type of ICC, or very small sample sizes (e.g., $N < 20$). The quality of evidence for the measures was moderate to high except for the AOCS (which was only supported by one very small study; $N = 23$). The best quality of evidence (high) was found for VTAS-R, VTAS-R-SF, and TPOCS-A.

3.2.6. Construct Validity

As with the self-report measures, we had studies on convergent validity in the form of comparisons with other alliance instruments, and predictive validity, that is, studies in which the alliance was used to predict psychotherapy outcome and/or dropout. Eight studies focused on five measurement instruments (VTAS-R-SF, ATAS, WAI—O, TPOCS-A, and AOCS) and explored convergent validity. Although most studies reported positive and significant correlations with other alliance measures, the quality of this evidence varied – primarily due to the small total N for some of the measures.

Predictive validity evidence was reported in nine studies for six instruments. In contrast to most of the other measurement properties, the evidence for predictive validity was much more varied. Only one study (on the AOCS) showed unequivocally positive outcome prediction. That study consisted of only 23 participants, so the quality of evidence was therefore rated as low. Neither the VTAS-R, the VTAS-R-SF, or the WAI-O reached the criterion of at least 75 % of the hypotheses supported. These measures were therefore rated as negative regarding predictive validity, with high quality evidence for VTAS-R, moderate for VTAS-R-SF, and low for the WAI—O. For ATAS and TPOCS-A, prediction of outcome was negative (i.e., non-significant or in the wrong direction), with the quality of the evidence rated as moderate.

4. Discussion

Our systematic review and COSMIN-based quality analysis revealed valuable insights into psychometric research on working alliance measurement instruments for adolescents. The Working Alliance Inventory-Short Form (WAI-S) was the most frequently used self-report measure, supported by high-quality evidence for structural validity (unidimensionality), internal consistency, and predictive validity. Other widely used tools included the WAI-SR and TASC/TASC-r, although the evidence supporting these measures was somewhat weaker and of lower quality. The only self-report instrument specifically developed for adolescents, the TAQS, was evaluated in one large study ($N = 679$ for structural validity and internal consistency, $N = 288$ for predictive validity and responsiveness) (Bickman et al., 2012), which provided high-quality evidence for positive structural validity and internal consistency. However, despite being studied with high-quality methods, the analyses did not support predictive validity and responsiveness for the TAQS.

Among the observer measures, the TPOCS-A, VTAS-R, and VTAS-R-SF stood out for their high-quality evidence supporting good internal consistency and reliability. The TPOCS-A also demonstrated positive concurrent validity, backed by high-quality evidence. However, observer measures faced challenges with predictive validity, showing inconsistent or negative results in their ability to predict therapeutic outcomes based on working alliance scores. None of the observer measures included evaluations of responsiveness, raising questions about their suitability for designs that involve repeated measurements of the alliance.

Many measures, both self-report and observer, fell short of demonstrating structural validity, raising concerns about whether they accurately capture the structure of the working alliance construct. To address this issue, large studies—ideally using confirmatory rather than exploratory factor models—are needed. Despite these limitations, most of the studies on structural validity showed a one-factor structure, particularly when evaluated using the COSMIN criteria for model fit. These criteria are somewhat less stringent than those commonly applied in SEM models (Kline, 2023), requiring only that *either* CFI > 0.95, RMSEA <0.06, *or* SRMR <0.08 (usually, *all* of these are required). Structural validity results also impact ratings of internal consistency; according to COSMIN, internal consistency should only be calculated for scales that demonstrate at least a low level of support for structural validity. Support for unidimensionality should be established when calculating alpha for the total scale score; for multidimensional scales, alpha should be calculated separately for subscales.

Additional limitations emerged in the assessment of test-retest reliability, which was adequately evaluated in only a few studies. A challenge in studying test-retest reliability for working alliances is identifying the optimal time interval. COSMIN requires that patients remain stable during the interim period, which, in the case of working alliance, we interpreted to mean that there should be no sessions

between measurements. The interval must be long enough to minimize recall bias yet short enough to ensure that patients' experiences of the alliance have not changed. The ideal trade-off between these two considerations may be to measure test-retest reliability once after a session and then again immediately before the next session. While this approach has been used in studies on adult populations (e.g., Kivity et al., 2022), no studies in our review employed this design.

Inter-rater reliability was the most frequently assessed property for observer measures; however, these studies often relied on very small sample sizes. In fact, only one study included a sample size of at least 100 participants (Hogue et al., 2006). Nevertheless, the combined sample sizes for the VTAS-R, VTAS-R-SF, and TPOCS-A exceeded 100, allowing these instruments to be rated as having high-quality evidence supporting inter-rater reliability. The psychometric property of measurement error was calculated in only two studies on self-report measures and was entirely absent in studies on observer measures.

This area represents an opportunity for the COSMIN framework to enhance methodological rigor in psychotherapy research. The COSMIN system advocates for calculating the minimally important change (MIC) for a given measure, employing an anchor-based approach to establish this threshold (de Vet et al., 2006). Researchers are then encouraged to test whether the smallest detectable change (SDC)—operationalized through indices such as the RCI (Jacobson & Truax, 1991)—falls below the MIC. This process ensures that the measurement instrument is sufficiently sensitive to detect changes that are clinically significant. Such an approach is particularly critical for repeated-measures studies of the therapeutic alliance, including time-lagged prediction models that investigate the dynamic interplay among the alliance, rupture-repair processes, and treatment outcomes. By ensuring that instruments are capable of capturing clinically meaningful changes, researchers can strengthen the validity of their findings while improving the interpretability of temporal associations.

Many self-report and observer measures failed to meet the COSMIN standard of supporting at least 75 % of hypotheses for predictive validity and responsiveness. On the one hand, this standard seems reasonable to prevent family-wise type I error, particularly in our case, as we relied on *p*-values rather than effect sizes, which COSMIN recommends. On the other hand, in a developing field like psychotherapy process research, our understanding of what outcomes to expect from variations in working alliance, including appropriate time lags, is still evolving. Consequently, it remains unclear whether the lack of predictive validity for a specific outcome reflects a flaw in the measure itself or simply that a relationship with alliance should not be expected for that particular outcome or time lag. Horvath (2018) even contends that outcome prediction is an inadequate indicator of construct validity for alliance measures, as many processes beyond alliance can predict outcomes. While this is a valid objection, we still believe that outcome prediction can serve as one indicator among several for construct validity, provided its limitations are considered.

In this regard, content validity is of paramount importance. If we are reasonably confident that the instrument effectively captures the working alliance construct without significant contamination from other processes, outcome prediction becomes a more convincing indicator of construct validity. However, our review found that content validity was often either overlooked or assumed to have been addressed when the adult version of the instrument was developed. Notably, research has yet to be conducted to assess the relevance, comprehensiveness, or comprehensibility of self-report questionnaire items from the perspective of adolescent patients. As a result, we cannot be certain that these questionnaires are understandable and relevant to adolescents' lived experiences in psychotherapy. According to COSMIN developers, content validity is regarded as the most critical measurement property (Mokkink et al., 2018).

Most alliance instrument developers appear to have followed the early alliance researchers (e.g., Alexander & Luborsky, 1986; Hartley & Strupp, 1983; Horvath & Greenberg, 1989) by using expert-based methods for test development. This approach contrasts with the COSMIN system's emphasis on qualitative interviews with patients and practitioners to assess a measure's relevance, comprehensiveness, and—specifically for patients—comprehensibility. Although the COSMIN system is likely the most comprehensive – perhaps even the only – framework for assessing the quality of measurement instruments, it was originally developed for patient-reported *outcome* measures. Therefore, adaptations are needed for its application to theory-based process measures like the working alliance. For example, we do not believe that assessing the content validity of observer-rated alliance measures requires qualitative interviews with patients, as patients may not be expected to recognize all aspects of the theoretical construct of working alliance in video-recorded sessions. However, qualitative interviews with therapists could still be a valuable method for evaluating content validity in observer-rated alliance measures.

4.1. Strengths and Limitations

This study has several strengths, including the use of the COSMIN system both for conducting the systematic search and evaluating the studies and instruments. The analyses were conducted by a large team of international experts, who are members of the TREATme research network, with all analyses and decisions regarding inclusion and exclusion based on assessments by at least two independent researchers. Another strength is our inclusive approach; we did not limit our review to studies that explicitly focused on psychometrics.

The large number of raters can also be seen as a limitation, with potentially more heterogenous perspectives on the ratings. However, we tried to counteract this by having one person check all ratings to ensure consistency. Another limitation of this review is the inclusion of only studies published in English, which may introduce language bias and exclude relevant findings from non-English sources. It is important to note that a review like this one represents a snapshot of the field at a specific moment. New research will alter the conclusions drawn in this study. Additionally, the absence of evidence is not equivalent to evidence of absence; we cannot conclude that measures less extensively studied or studied with low-quality evidence are inferior to those with high-quality evidence. However, we can be more confident about the measurement properties of instruments with robust supporting evidence.

4.2. Recommendations for Applied Alliance Research

Based on our review, we recommend using the WAI-S or the TAQS when a self-report instrument is needed for assessing working alliance in adolescents. This recommendation is primarily due to these instruments' strong support for structural validity and internal consistency. The WAI-S has broader support, with more studies backing it and better evidence for predictive validity. However, the TAQS may be more developmentally appropriate, as it was specifically designed with this age group in mind. Among the observer measures, the TPOCS-A currently offers the best support in terms of internal consistency, reliability, and concurrent validity. At present, however, no observer measure has demonstrated high-quality evidence for predictive validity.

4.3. Recommendations for Psychometric Research on Alliance Instruments

Our study highlights the need for improvements, particularly in the study of content validity of measurement instruments. Many tools have been adapted from adult-focused instruments and may lack sensitivity to the developmental levels of adolescents. Given the vulnerabilities of this age group, the experience, manifestation, and role of the alliance in therapy may differ from those in adult psychotherapy. Therefore, developers of working alliance instruments should take the developmental phase into account.

None of the studies supported the theoretical three-factor structure adapted from instruments developed for adults, suggesting that separating goals/tasks from the emotional bond may be less relevant for adolescents. This again brings up questions about adapting adult-focused instruments to adolescents. Comprehensibility may be particularly important, especially for self-report instruments used with younger adolescents. To address this gap, future measures should be developed with more significant input from adolescents, ensuring that item relevance, language, and clarity align with their experiences in therapy. Our review also emphasizes the need for expanded cross-cultural testing, as few existing measures have demonstrated robustness across diverse cultural contexts, which is essential for broad applicability.

Additionally, the responsiveness of working alliance measures — their ability to accurately track changes over time—remains underexplored. Future studies should prioritize this property to enhance the utility of these instruments in longitudinal research. This is especially important given the field's shift toward session-by-session prediction models (Falkenström, 2024; Zilcha-Mano & Fisher, 2022). Future reviews could also consider pooling studies to enable more sophisticated analyses of psychometric properties.

4.4. Conclusions

In conclusion, this systematic review reveals progress, but there are also significant gaps in the psychometric evaluation of working alliance measures in adolescents. While certain self-report and observer instruments, such as the WAI-S and TPOCS-A, support fundamental properties like internal consistency and construct validity, most instruments fall short in essential areas, particularly content validity, responsiveness, and cross-cultural applicability. Our findings underscore the need for more robust and developmentally sensitive measures for adolescent populations. Future research would benefit from focusing on these underexplored areas, integrating patient and therapist perspectives to enhance the relevance and applicability of these instruments. By addressing these limitations, the field can develop reliable tools that effectively capture the working alliance construct and support the growing emphasis on predictive, session-by-session analyses in therapeutic contexts.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT and Grammarly Pro for proofreading. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We gratefully acknowledge the help of research librarians Trine Lacoppidan Kæstel and Vibeke Rabjerg Grünbaum, Psychiatric Research Unit, Slagelse in the construction of search strings, updates and reference management. The review was conducted by researchers involved in the COST Action on European Network on Individualized Psychotherapy Treatment of Young People with Mental Disorders (TREATme; www.cost.eu/actions/CA16102/) funded by the European Cooperation in Science and Technology (COST). We extend our thanks to the full COST TREATme network for continuous support and encouragement.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cpr.2025.102586.

References

Alexander, L. B., & Luborsky, L. (1986). The Penn Helping Alliance Scales. In *The psychotherapeutic process: A research handbook* (pp. 325–366). Guilford Press.

Alvarez, I., Herrero, M., Martínez-Pampliega, A., & Escudero, V. (2021). Measuring perceptions of the therapeutic alliance in individual, family, and group therapy from a systemic perspective: Structural validity of the SOFTA-s. *Family Process, 60*(2), 302–315. https://doi.org/10.1111/famp.12565

Anderson, S. R., Johnson, L. N., Witting, A. B., Miller, R. B., Bradford, A. B., Hunt, Q. A., & Bean, R. A. (2024). Validation of the intersession alliance measure: Individual, couple, and family versions. *Journal of Marital and Family Therapy*. https://doi.org/10.1111/jmft.12702

Anderson, R. E. E., Spence, S. H., Donovan, C. L., March, S., Prosser, S., & Kenardy, J. (2012). Working alliance in online cognitive behavior therapy for anxiety disorders in youth: Comparison with clinic delivery and its role in predicting outcome. *Journal of Medical Internet Research, 14*(3), 86–101. https://doi.org/10.2196/jmir.1848

Ayotte, M.-H., Lanctot, N., & Tourigny, M. (2016). How the working alliance with adolescent girls in residential care predicts the trajectories of their behavior problems. *Residential Treatment for Children & Youth, 33*(2), 135–154. https://doi.org/10.1080/0886571X.2016.1175994

Ayotte, M.-H., Lanctot, N., & Tourigny, M. (2017). The association between the working alliance with adolescent girls in residential care and their trauma-related symptoms in emerging adulthood. *Child & Youth Care Forum, 46*(4), 601–620. https://doi.org/10.1007/s10566-017-9398-x

van Benthem, P., Spijkerman, R., Blanken, P., Kleinjan, M., Vermeiren, R., & Hendriks, V. M. (2020). A dual perspective on first-session therapeutic alliance: Strong predictor of youth mental health and addiction treatment outcome. *European Child and Adolescent Psychiatry, 29*(11), 1593–1601. https://doi.org/10.1007/s00787-020-01503-w

van Benthem, P., van der Lans, R. M., Lamers, A., Blanken, P., Spijkerman, R., Vermeiren, R., & Hendriks, V. M. (2024). The Working Alliance Inventory - short version: Psychometric properties of the patient and therapist form in youth mental health and addiction care. *BMC Psychol, 12*(1), 319. https://doi.org/10.1186/s40359-024-01754-1

Bickman, L., Athay, M., Riemer, M., Lambert, E., Kelley, S., Breda, C., & Vides de Andrade, A. (2010). *Manual of the Peabody Treatment Progress Battery*. Nashville, TN: Vanderbilt University.

Bickman, L., de Andrade, A. R. V., Athay, M. M., Chen, J. I., De Nadai, A. S., Jordan-Arthur, B. L., & Karver, M. S. (2012). The relationship between change in therapeutic alliance ratings and improvement in youth symptom severity: Whose ratings matter the most? *Administration and Policy in Mental Health and Mental Health Services Research, 39*(1–2), 78–89. https://doi.org/10.1007/s10488-011-0398-0

Blais, M. A. (2004). Development of an inpatient treatment alliance scale. *The Journal of Nervous and Mental Disease, 192*(7). https://doi.org/10.1097/01.nmd.0000131911.53489.af

Bleiberg, E. (2013). Mentalizing-based treatment with adolescents and families. *Child and Adolescent Psychiatric Clinics of North America, 22*(2), 295–330. https://doi.org/10.1016/j.chc.2013.01.001

Bordin, E. S. (1979). The generalizability of the psychoanalytic concept of the working alliance. *Psychotherapy: Theory, Research & Practice, 16*(3), 252–260. https://doi.org/10.1037/h0085885

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*(4), 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061

Cirasola, A., & Midgley, N. (2023). The alliance with young people: Where have we been, where are we going? *Psychotherapy, 60*(1), 110–118. https://doi.org/10.1037/pst0000461

Cirasola, A., Midgley, N., Fonagy, P., Consortium, I, & Martin, P. (2022). The therapeutic alliance in psychotherapy for adolescent depression: Differences between treatment types and change over time. *Journal of Psychotherapy Integration, 32*(3). https://doi.org/10.1037/int0000264. No-Specified.

Cirasola, A., Midgley, N., Fonagy, P., Impact Consortium, & Martin, P. (2021). The alliance-outcome association in the treatment of adolescent depression. *Psychotherapy, 58*(1), 95–108. https://doi.org/10.1037/pst0000366

Cirasola, A., Midgley, N., Fonagy, P., & Martin, P. (2021). The factor structure of the Working Alliance Inventory short-form in youth psychotherapy: An empirical investigation. *Psychotherapy Research, 31*(4), 535–547. https://doi.org/10.1080/10503307.2020.1765041

Cooper, M., Connor, C., Orloff, N., Herrington, J. D., & Timko, C. A. (2024). Therapeutic alliance in family-based treatment of anorexia nervosa: In-person versus telehealth. *Clinical Psychology & Psychotherapy, 31*(3), Article e3017. https://doi.org/10.1002/cpp.3017

Creed, T. A., & Kendall, P. C. (2005). *Therapeutic Alliance Scales for Children–Revised (TASC-R)*. APA PsycTests.

Crits-Christoph, P., Connolly Gibbons, M. B., & Mukherjee, D. (2013). Psychotherapy process-outcome research. In M. J. Lambert (Ed.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (6th ed., pp. 298–340). John Wiley & Sons.

Cronbach, L., & Meehl, P. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302. https://doi.org/10.1037/h0040957

Darchuk, A., Wang, V., David, W., Fende, J., Anderson, T., & Horvath, A. (2000). *Manual for the Working Alliance Inventory – Observer form (WAI-O): Revision IV*. Ohio University.

DeRubeis, R. J., Brotman, M. A., & Gibbons, C. J. (2005). A conceptual and methodological analysis of the nonspecifics argument. *Clinical Psychology: Science and Practice, 12*, 174–183. https://doi.org/10.1093/clipsy/bpi022

Diamond, G., Ruan-Iu, L., Winston-Lindeboom, P., Rivers, A. S., Weissinger, G., & Roeske, M. (2024). Treatment readiness in psychiatric residential care for adolescents. *Administration and Policy in Mental Health and Mental Health Services Research, 51*(6), 877–888. https://doi.org/10.1007/s10488-024-01393-z

Duncan, B. L., Miller, S. D., Sparks, J. A., Claud, D. A., Beach, P., Reynolds, L. R., & Johnson, L. D. (2003). The Session Rating Scale: Preliminary psychometric properties of a working alliance measure. *Journal of Brief Therapy, 3*, 3–12.

Ekeblad, A., Falkenström, F., & Holmqvist, R. (2016). Reflective functioning as predictor of working alliance and outcome in the treatment of depression. *Journal of Consulting and Clinical Psychology, 84*(1), 67–78. https://doi.org/10.1037/ccp0000055

Elvins, R., & Green, J. (2008). The conceptualization and measurement of therapeutic alliance: An empirical review. *Clinical Psychology Review, 28*, 1167–1187. https://doi.org/10.1016/j.cpr.2008.04.002

Escudero, V., Friedlander, M. L., Kivlighan, D. M., Orlowski, E., & Abascal, A. (2022). Mapping the progress of the process: Codevelopment of the therapeutic alliance with maltreated adolescents. *Journal of Counseling Psychology, 69*(5), 656–666. https://doi.org/10.1037/cou0000621

Falkenström, F. (2024). Time-lagged panel models in psychotherapy process and mechanisms of change research: Methodological challenges and advances. *Clinical Psychology Review, 110*, Article 102435. https://doi.org/10.1016/j.cpr.2024.102435

Falkenström, F., Hatcher, R. L., Skjulsvik, T., Larsson, M. H., & Holmqvist, R. (2015). Development and validation of a 6-item working alliance questionnaire for repeated administrations during psychotherapy. *Psychological Assessment, 27*(1), 169–183. https://doi.org/10.1037/pas0000038

Faw, L., Hogue, A., Johnson, S., Diamond, G. M., & Liddle, H. A. (2005). The Adolescent Therapeutic Alliance Scale (ATAS): Initial psychometrics and prediction of outcome in family-based substance abuse prevention counseling. *Psychotherapy Research, 15*(1–2), 141–154. https://doi.org/10.1080/10503300512331326994

Flückiger, C., Del Re, A. C., Wampold, B. E., & Horvath, A. O. (2018). The alliance in adult psychotherapy: A meta-analytic synthesis. *Psychotherapy, 55*(4), 316–340. https://doi.org/10.1037/pst0000172

Flückiger, C., Rubel, J., Del Re, A. C., Horvath, A. O., Wampold, B. E., Crits-Christoph, P., … R.,…Barber, J. P.. (2020). The reciprocal relationship between alliance and early treatment symptoms: A two-stage individual participant data meta-analysis. *Journal of Consulting and Clinical Psychology, 88*(9), 829–843. https://doi.org/10.1037/ccp0000594

Freud, S. (1949). *An outline of psychoanalysis*. W. W. Norton.

Friedlander, M. L., Escudero, V., & Heatherington, L. (2006). Introducing the system for observing family therapy alliances. In *Therapeutic alliances in couple and family therapy: An empirically informed guide to practice* (pp. 31–49). American Psychological Association. https://doi.org/10.1037/11410-002.

Gergov, V., Marttunen, M., Lindberg, N., Lipsanen, J., & Lahti, J. (2021). Therapeutic Alliance: A comparison study between adolescent patients and their therapists. *International Journal of Environmental Research and Public Health, 18*(21). https://doi.org/10.3390/ijerph182111238

Greenson, R. R. (1965). The working alliance and the transference neurosis. *The Psychoanalytic Quarterly, 34*, 155–179.

de Haan, A. M., Boon, A. E., de Jong, J. T. V. M., Hoeve, M., & Vermeiren, R. R. J. M. (2013). A meta-analytic review on treatment dropout in child and adolescent outpatient mental health care. *Clinical Psychology Review, 33*(5), 698–711. https://doi.org/10.1016/j.cpr.2013.04.005

Haggerty, G., Siefert, C. J., Sinclair, S. J., Zodan, J., Babalola, R., & Blais, M. A. (2015). Validation of a measure of alliance for an adolescent inpatient setting. *Clinical Psychology & Psychotherapy, 22*(4), 357–363. https://doi.org/10.1002/cpp.1901

Hartley, D. E., & Strupp, H. H. (1983). The therapeutic alliance: Its relationship to outcome in brief psychotherapy. In J. Masling (Ed.), *Vol. 1. Empirical studies of psychoanalytical theories* (pp. 1–37). Analytic Press.

Hatcher, R. L., & Gillaspy, J. A. (2006). Development and validation of a revised short version of the Working Alliance Inventory. *Psychotherapy Research, 16*, 12–25. https://doi.org/10.1080/10503300500352500

Hauber, K., & Boon, A. (2022). First-session therapeutic relationship and outcome in high risk adolescents intensive group psychotherapeutic programme. *Frontiers in Psychology, 13*, Article 916888. https://doi.org/10.3389/fpsyg.2022.916888

Hauber, K., Boon, A., & Vermeiren, R. (2020). Therapeutic relationship and dropout in high-risk adolescents' intensive group psychotherapeutic programme. *Frontiers in Psychology, 11*. https://doi.org/10.3389/fpsyg.2020.533903

Hawley, K. M., & Garland, A. F. (2008). Working alliance in adolescent outpatient therapy: Youth, parent and therapist reports and associations with therapy outcomes. *Child & Youth Care Forum, 37*(2), 59–74. https://doi.org/10.1007/s10566-008-9050-x

Hogue, A., Dauber, S., Stambaugh, L. F., Cecero, J. J., & Liddle, H. A. (2006). Early therapeutic alliance and treatment outcome in individual and family therapy for adolescent behavior problems. *Journal of Consulting and Clinical Psychology, 74*(1), 121–129. https://doi.org/10.1037/0022-006X.74.1.121

Horvath, A. O. (2018). Research on the alliance: Knowledge in search of a theory. *Psychotherapy Research, 28*(4), 499–516. https://doi.org/10.1080/10503307.2017.1373204

Horvath, A. O., & Greenberg, L. S. (1989). Development and validation of the Working Alliance Inventory. *Journal of Counseling Psychology, 36*, 223–233.

Hou, Y., Hu, J., Zhang, X., Zhao, J., Yang, X., Sun, X., … Fang, L. (2024). Validation of the capacity for the psychotherapy process scale for use in adolescent patients. *Res Child Adolesc Psychopathol.*. https://doi.org/10.1007/s10802-024-01209-6

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59*, 12–19.

Jacoby, R. J., Smilansky, H., Shin, J., Wu, M. S., Small, B. J., Wilhelm, S., … Geller, D. A. (2021). Longitudinal trajectory and predictors of change in family accommodation during exposure therapy for pediatric OCD. *Journal of Anxiety Disorders, 83*. https://doi.org/10.1016/j.janxdis.2021.102463

Jewell, T., Herle, M., Serpell, L., Eivors, A., Simic, M., Fonagy, P., & Eisler, I. (2023). Attachment and mentalization as predictors of outcome in family therapy for adolescent anorexia nervosa. *European Child and Adolescent Psychiatry, 32*(7), 1241–1251. https://doi.org/10.1007/s00787-021-01930-3

Johnson, L. N., Ketring, S. A., & Anderson, S. R. (2013). Confirmatory factor analysis of the Family Therapy Alliance Scale: An evaluation of factor structure across parents and adolescents referred for at risk services. *Contemporary Family Therapy: An International Journal, 35*(1), 121–136. https://doi.org/10.1007/s10591-012-9221-7

Johnson, S., Hogue, A., Diamond, G., Leckrone, J., & Liddle, H. A. (1998). *Scoring manual for the Adolescent Therapeutic Alliance Scale (ATAS)*. Temple University.

Kang, E., Gioia, A., Pugliese, C. E., Islam, N. Y., Martinez-Pedraza, F., Girard, R. M., … Lerner, M. D. (2021). Alliance-outcome associations in a community-based social skills intervention for youth with autism spectrum disorder. *Behavior Therapy, 52*(2), 324–337. https://doi.org/10.1016/j.beth.2020.04.006

Karver, M., Shirk, S. R., Day, R., Field, S., & Handelsman, J. B. (2003). *Rater's manual for the Alliance Observation Coding System*.

Karver, M., Shirk, S. R., Handelsman, J. B., Fields, S., Crisp, H., Gudmundsen, G., & McMakin, D. (2008). Relationship processes in youth psychotherapy: Measuring alliance, alliance-building behaviors, and client involvement. *Journal of Emotional and Behavioral Disorders, 16*(1), 15–28. https://doi.org/10.1177/1063426607312536

Karver, M. S., De Nadai, A. S., Monahan, M., & Shirk, S. R. (2018). Meta-analysis of the prospective relation between alliance and outcome in child and adolescent psychotherapy. *Psychotherapy, 55*(4), 341–355. https://doi.org/10.1037/pst0000176

Kermarrec, S., Kabuth, B., Bursztejn, C., & Guillemin, F. (2006). French adaptation and validation of the Helping Alliance Questionnaires for Child, Parents, and Therapist. *The Canadian Journal of Psychiatry, 51*(14), 913–922. https://doi.org/10.1177/070674370605101407

King, S. C., Richner, K. A., Tuliao, A. P., Kennedy, J. L., & McChargue, D. E. (2019). A comparison between telehealth and face-to-face delivery of a brief alcohol intervention for college students. *Substance Abuse, 41*(4), 501–509. https://doi.org/10.1080/08897077.2019.1675116

Kivity, Y., Strauss, A. Y., Elizur, J., Weiss, M., Cohen, L., & Huppert, J. D. (2022). Patterns of alliance development in cognitive behavioral therapy versus attention bias modification for social anxiety disorder: Sawtooth patterns and sudden gains. *Journal of Clinical Psychology, 78*(2), 122–136. https://doi.org/10.1002/jclp.23219

Kivlighan, D. M., Jr., Escudero, V., Friedlander, M. L., & Orlowski, E. (2022). Illustrating systemic change in family therapy: How therapists' and clients' alliance perceptions codevelop over time. *Psychotherapy Research, 1-12*. https://doi.org/10.1080/10503307.2022.2071131

Kline, R. B. (2023). *Principles and practice of structural equation modeling* (5 ed.). Guilford Press.

Koocher, G. P. (2003). Ethical issues in psychotherapy with adolescents. *Journal of Clinical Psychology, 59*(11), 1247–1256. https://doi.org/10.1002/jclp.10215

Lindqvist, K., Mechler, J., Falkenström, F., Carlbring, P., Andersson, G., & Philips, B. (2023). Therapeutic alliance is calming and curing – The interplay between alliance and emotion regulation as predictors of outcome in internet-based treatments for adolescent depression. *Journal of Consulting and Clinical Psychology, 91*(7), 426–437. https://doi.org/10.1037/ccp0000815

McLeod, B. D., & Weisz, J. R. (2005). The Therapy Process Observational Coding System-Alliance scale: Measure characteristics and prediction of outcome in usual clinical practice. *Journal of Consulting and Clinical Psychology, 73*(2), 323–333. https://doi.org/10.1037/0022-006X.73.2.323

McLeod, J., Stänicke, E., Oddli, H. W., Smith, S., Pearce, P., & Cooper, M. (2024). How do we know whether treatment has failed? Paradoxical outcomes in counseling with young people. *Frontiers in Psychology, 15*, 1390579. https://doi.org/10.3389/fpsyg.2024.1390579

Meeks, J. E., & Bernet, W. (2001). *The fragile alliance: An orientation to the outpatient psychotherapy of the adolescent* (5 ed.). Krieger.

Mekori-Domachevsky, E., Matalon, N., Mayer, Y., Shiffman, N., Lurie, I., Gothelf, D., & Dekel, I. (2023). Internalizing symptoms impede adolescents' ability to transition from in-person to online mental health services during the 2019 coronavirus disease pandemic. *Journal of Telemedicine and Telecare, 29*(9), 725–730. https://doi.org/10.1177/1357633X211021293

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Bmj, 339*, Article b2535. https://doi.org/10.1136/bmj.b2535

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., … 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, 19*(4), 539–549. https://doi.org/10.1007/s11136-010-9606-8

Mokkink, L. B., 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, 27*(5), 1171–1179. https://doi.org/10.1007/s11136-017-1765-4

Nissling, L., Weineland, S., Vernmark, K., Radvogin, E., Engstroem, A. K., Schmidt, S., … Hursti, T. (2023). Effectiveness of and processes related to internet-delivered acceptance and commitment therapy for adolescents with anxiety disorders: A randomized controlled trial. *Research in Psychotherapy - Psychopathology Process and Outcome, 26*(2). https://doi.org/10.4081/ripppo.2023.681

Ormhaug, S. M., Shirk, S. R., & Wentzel-Larsen, T. (2015). Therapist and client perspectives on the alliance in the treatment of traumatized adolescents. *European Journal of Psychotraumatology, 31*(6). https://doi.org/10.3402/ejpt.v6.27705

Ovenstad, K. S., Jensen, T. K., & Ormhaug, S. M. (2022). Four perspectives on traumatized youths' therapeutic alliance: Correspondence and outcome predictions. *Psychotherapy Research, 32*(6), 820–832. https://doi.org/10.1080/10503307.2021.2011983

Ovenstad, K. S., Ormhaug, S. M., & Jensen, T. K. (2023). The relationship between youth involvement, alliance and outcome in trauma-focused cognitive behavioral therapy. *Psychotherapy Research, 33*(3), 316–327. https://doi.org/10.1080/10503307.2022.2123719

Ovenstad, K. S., Ormhaug, S. M., Shirk, S. R., & Jensen, T. K. (2020). Therapists' behaviors and youths' therapeutic alliance during trauma-focused cognitive behavioral therapy. *Journal of Consulting and Clinical Psychology, 88*(4), 350–361. https://doi.org/10.1037/ccp0000465

Piaget, J. (1977). *The development of thought: Equilibration of cognitive structures*. Viking.

Pinsof, W. M., & Catherall, D. R. (1986). The integrative psychotherapy alliance: Family, couple, and individual therapy scales. *Journal of Marital and Family Therapy, 12*(2), 137–151. https://doi.org/10.1111/j.1752-0606.1986.tb01631.x

Pinsof, W. M., Zinbarg, R., & Knobloch-Fedders, L. M. (2008). Factorial and construct validity of the revised short form integrative psychotherapy alliance scales for family, couple, and individual therapy. *Family Process, 47*(3), 281–301. https://doi.org/10.1111/j.1545-5300.2008.00254.x

Prinsen, C. A. C., Mokkink, L. B., Bouter, L. M., Alonso, J., Patrick, D. L., de Vet, H. C. W., & Terwee, C. B. (2018). COSMIN guideline for systematic reviews of patient-reported outcome measures. *Quality of Life Research, 27*(5), 1147–1157. https://doi.org/10.1007/s11136-018-1798-3

Puls, H. C., Schmidt, R., & Hilbert, A. (2019). Therapist adherence and therapeutic alliance in individual cognitive-behavioural therapy for adolescent binge-eating disorder. *European Eating Disorders Review, 27*(2), 182–194. https://doi.org/10.1002/erv.2650

Rienecke, R. D., Richmond, R., & Lebow, J. (2016). Therapeutic alliance, expressed emotion, and treatment outcome for anorexia nervosa in a family-based partial hospitalization program. *Eating Behaviors, 22*, 124–128. https://doi.org/10.1016/j.eatbeh.2016.06.017

Rollin, M., Pascuzzo, K., & Lanctot, N. (2024). Do adolescent girls' relationships with their parents influence their perceptions of the therapeutic alliance and group climate in residential care? *Child & Family Social Work, 29*(1), 205–216. https://doi.org/10.1111/cfs.13065

Shelef, K., & Diamond, G. M. (2008). Short form of the Revised Vanderbilt Therapeutic Alliance Scale: Development, reliability, and validity. *Psychotherapy Research, 18*(4), 433–443. https://doi.org/10.1080/10503300701810801

Shirk, S. R., & Saiz, C. C. (1992). *Therapeutic Alliance scales for children (TASC)*.

Silvers, J. A. (2022). Adolescence as a pivotal period for emotion regulation development. *Current Opinion in Psychology, 44*, 258–263. https://doi.org/10.1016/j.copsyc.2021.09.023

Speranza, M., & Midgley, N. (2017). Profile of mental functioning for adolescents (MA Axis). In V. Lingiardi, & N. McWilliams (Eds.), *Psychodynamic diagnostic manual: PDM-2* (pp. 263–322). Guilford Press.

Steil, R., Weiss, J., Renneberg, B., Gutermann, J., & Rosner, R. (2023). Effect of therapeutic competence, adherence, and alliance on treatment outcome in youth with PTSD treated with developmentally adapted cognitive processing therapy. *Child Abuse & Neglect, 141*, Article 106221. https://doi.org/10.1016/j.chiabu.2023.106221

Steinberg, L., Dahl, R., Keating, D., Kupfer, D. J., Masten, A. S., & Pine, D. S. (2015). The study of developmental psychopathology in adolescence: Integrating affective neuroscience with the study of context. In D. Cicchetti, & D. J. Cohen (Eds.), *Developmental psychopathology* (pp. 710–741). https://doi.org/10.1002/9780470939390.ch18

Sterba, R. (1934). The fate of the ego in analytic therapy. *The International Journal of Psychoanalysis, 15*, 117–126.

Terwee, C. B., Jansma, E. P., Riphagen, I. I., & de Vet, H. C. W. (2009). Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. *Quality of Life Research, 18*(8), 1115–1123. https://doi.org/10.1007/s11136-009-9528-5

Terwee, C. B., Mokkink, L. B., Knol, D. L., Ostelo, R. W. J. G., Bouter, L. M., & de Vet, H. C. W. (2012). Rating the methodological quality in systematic reviews of studies on measurement properties: A scoring system for the COSMIN checklist. *Quality of Life Research, 21*(4), 651–657. https://doi.org/10.1007/s11136-011-9960-1

Terwee, C. B., Prinsen, C. A. C., Chiarotto, A., Westerman, M. J., Patrick, D. L., Alonso, J., … Mokkink, L. B. (2018). COSMIN methodology for evaluating the content validity of patient-reported outcome measures: A Delphi study. *Quality of Life Research, 27*(5), 1159–1170. https://doi.org/10.1007/s11136-018-1829-0

Tracey, T. J., & Kokotovic, A. M. (1989). Factor structure of the Working Alliance Inventory. *Psychological Assessment, 1*, 207–210. https://doi.org/10.1037/1040-3590.1.3.207

Usami, S., Murayama, K., & Hamaker, E. L. (2019). A unified framework of longitudinal models to examine reciprocal relations. *Psychological Methods, 24*, 637–657. https://doi.org/10.1037/met0000210

VandenBos, G. R. (Ed.). (2015). *APA dictionary of psychology, 2nd ed*:https://doi.org/10.1037/14646-000]. American Psychological Association. doi: https://doi.org/10.1037/14646-000.

de Vet, H. C., Terwee, C. B., Ostelo, R. W., Beckerman, H., Knol, D. L., & Bouter, L. M. (2006). Minimal changes in health status questionnaires: Distinction between minimally detectable change and minimally important change. *Health and Quality of Life Outcomes, 4*(1), 54. https://doi.org/10.1186/1477-7525-4-54

Warschburger, P., Gmeiner, M. S., Bondü, R., Klein, A. M., Busching, R., & Elsner, B. (2023). Self-regulation as a resource for coping with developmental challenges during middle childhood and adolescence: The prospective longitudinal PIERYOUTH-study. *BMC Psychology, 11*(1), 97. https://doi.org/10.1186/s40359-023-01140-3

Webb, C. A., DeRubeis, R. J., Amsterdam, J. D., Shelton, R. C., Hollon, S. D., & Dimidjian, S. (2011). Two aspects of the therapeutic alliance: Differential relations with depressive symptom change. *Journal of Consulting and Clinical Psychology, 79*, 279–283. https://doi.org/10.1037/a0023252

Zilcha-Mano, S. (2017). Is the alliance really therapeutic? Revisiting this question in light of recent methodological advances. *American Psychologist, 72*(4), 311–325. https://doi.org/10.1037/a0040435

Zilcha-Mano, S., & Fisher, H. (2022). Distinct roles of state-like and trait-like patient–therapist alliance in psychotherapy. *Nature Reviews Psychology, 1*(4), 194–210. https://doi.org/10.1038/s44159-022-00029-z