**Interrater reliability of Criterion A of the Alternative Model for Personality Disorder (DSM 5 - Section III): A Meta-analysis**

Samantha Young & Peter Beazley

Department of Clinical Psychology and Psychological Therapies

Norwich Medical School

University of East Anglia

**© 2023, American Psychological Association. This paper is not the copy of record and may not exactly replicate the final, authoritative version of the article. Please do not copy or cite without authors' permission. The final article will be available, upon publication, via its DOI: 10.1037/per0000631**

This review did not receive a grant or funding and no conflicts of interest were reported.

# Abstract

The Alternative Model of Personality Disorder (AMPD) is currently included in Section III of the Diagnostic and Statistical Manual for Mental Disorders – Fifth Edition (DSM-5). This review sought to summarise the literature concerning interrater reliability (IRR) of the AMPD. Despite high heterogeneity, meta-analysis provided tentative support for the IRR of Criterion A of the AMPD, with pooled ICCs for the Level of Personality Functioning Scale (LPFS) and its domains falling above DSM acceptability levels. Sub-group analysis of the LPFS suggested IRR scores could be improved by using a specific AMPD Structured Clinical Interview (SCI). Further research should in particular consider the IRR of Criterion B elements of the AMPD and overall PD diagnosis, where insufficient data was available to draw conclusions in the present study.

Keywords: *Personality Disorder, interrater reliability, IRR, Alternative Model of Personality Disorder, AMPD, Levels of Functioning, LFPS*

# Introduction

Since their development, both of the taxonomic systems of the Diagnostic and Statistical Manual for Mental Disorders (DSM) and the International Classification of Diseases (ICD) have primarily understood Personality Disorder as a concept best described through a collection of discrete, categorical, diagnoses. However, over the last 15 years, there has been a particular interest in developing dimensional approaches to the assessment of Personality Disorder (Widiger et al., 2009), with such an approach being reflected both in the DSM-5’s ‘Alternative Model’ of Personality Disorder (AMPD; DSM-5, Section III [American Psychiatric Association, 2013]) and the ICD-11’s arguably more radical decision to abolish all specific categories of Personality Disorder (World Health Organisation, 2018).

Dimensional approaches to personality classification are, of course, not new. Outside the area of psychopathology, such approaches have become widely acceped (Costa & Widiger, 1994; Eysenck & Eysenck, 1975). In the clinical context, such approaches are attractive as they appear to have the potential to provide at least a partial solution to some of the difficulties associated with categorical approaches, specifically the overuse of the “PD not otherwise specified” (PDNOS) diagnosis (Verheul & Widiger, 2004), the frequent need for multiple categorical diagnoses to accurately describe a particular presentation, and the high heterogeneity of symptoms that appear within diagnostic categories (Widiger et al., 2009; Wright & Zimmerman, 2015). Yet, the task of developing a dimensional approach that demonstrates reliability, validity and acceptability in clinical practice has not been without difficulty. Indeed, as the DSM-5 was launched, concerns were raised around the AMPD’s clinical utility, and the model was deemed too complex and theory laden for clinical practice (Clarkin & Huprich, 2011; Pincus, 2011; Pilkonis et al., 2011), ultimately leading to its inclusion in ‘Section III’ of the DSM-5.

Regardless of the language used to describe personality characteristics, a clear issue for any taxonomic system is the need to identify the point at which a characteristic, trait or disorder is considered maladaptive or dysfunctional. One particular challenge is that perceptions of what constitutes “maladaptive” vary depending on the cultural context; PDs are diagnosed less frequently in collectivistic cultures (Winsper et al., 2019), and in certain cultures, particular personality presentations have not been described at all (for instance, Borderline PD has never been included in the Chinese Classification of Mental Disorders (CCMD-3; Chinese Psychiatry Association, 2000)). Yet, quantifying the level of impairment associated with personality functioning seems an important goal, with research suggesting that such impairment is the largest predictor of both treatment outcome and distress (Crawford et al., 2011; Bender et al., 2011). A structured assessment of the extent of dysfunction has the potential to provide a more clinically useful focus for diagnosis and prevent unhelpful behavioral descriptions of PD (Mulder, 2021).

Criterion A of the AMPD specifies the requirement for impairment in personality functioning, operationalizing this through the Level of Personality Functioning Scale (LPFS; Bender et al., 2011). The dimensional constructs are categorised into “Self” (including *identity* and *self-direction*) and “Interpersonal” (including *intimacy* and *empathy*) domains which are rated on a five-point scale. Moderate impairment or greater across at least two domains meets clinical threshold. Criterion B identifies pathological personality traits. Twenty-five trait facets are identified under the following categories: Negative Affectivity, Detachment, Antagonism, Disinhibition and Psychoticism. The facets are rated on a four-level scale of descriptivism.

Measurement of functional impairment in the ICD-11 shares some similarities with the approach adopted by the DSM-5 (Mulder, 2021; Hemmati et al., 2021), with assessment occurring across three levels of severity (mild, moderate and severe) and five personality trait domains; Negative Affectivity, Detachment, Disinhibition, Dissociality, and Anankastia (WHO, 2018). Given the broad similarities in approaches, it is hoped that the growing body of research surrounding the AMPD will be generalisable to the ICD-11 also (Mulder, 2021; Mulder & Tyrer, 2019).

**Interrater Reliability**

Interrater Reliability (IRR) is a fundamental assessment tool for diagnostic accuracy (Kraemer et al., 2012). It enables researchers to quantify agreement between two or more raters and thus measure the consistency with which diagnostic criteria are applied. IRR is of particular interest in PD diagnosis where within-category heterogeneity and multiple diagnoses are common (Widiger et al., 2009). Arguably, diagnostic systems with higher IRR indicate a shared conceptual understanding of the diagnosis and its application and are less prone to errors caused by human judgment, inexperience, or heuristics (Aboraya et al., 2006; Tversky & Kahneman, 1974).

A previous review conducted by Samuel (2015) found overall moderate levels of IRR for the categorical system of PD. They reported a median kappa value for the diagnosis of any PD of .52 and an individual PD of .40. Their findings suggested that using a structured clinical assessment (the Shedler-Westen Assessment Procedure; Westen & Shedler, 1999) produced increased levels of IRR (*k*=.61), however variations in study design had little impact on the overall score (Samuel, 2015).

**Aims**

This review builds on Samuel’s (2015) review concerning the IRR of PD diagnoses in the context of the AMPD and specifically Criterion A and the associated LPFS. To the authors’ knowledge, no systematic review or meta-analysis exists summarising literature on IRR of the LPFS and its associated instruments. Therefore, this review aims to effectively summarise existing studies which report a statistical measure of IRR for the LPFS.

At the point of pre-registration of the present review, it was hoped that a systematic review of IRR in relation to all elements of the AMPD could be conducted, however insufficient published research in relation to other elements including Criterion B and the overall diagnosis was available. The manuscript therefore presents the findings in relation to Criterion A only, which is considered in terms of a meta-analysis. Outcomes are reported in relation to DSM-5 acceptability thresholds for IRR (>.40; Kraemer et al., 2012) and reporting guideline thresholds for Intraclass Correlation Coefficients (ICCs); a dimensional measure of IRR (Koo & Li, 2016).

**Method**

**Protocol and Registration**

The review protocol was pre-registered on the PROSPERO international prospective register of systematic reviews (PROSPERO, reference CRD42021254416).

**Search Strategy**

A systematic search of published literature was conducted using the electronic databases PsycINFO, MEDLINE, CINAHL, and Scopus. Search terms "personality disorder\*" AND ("alternative model of personality disorder\*" or "ampd") were used. Broad search terms were used due to the limited research available. Reference lists and citations of key review papers and accepted papers were hand searched. Searches were conducted on 4th June 2021.

# Eligibility

 Studies were only considered in the review if they met all of the following inclusion criteria: (a) Empirical research studies which applied one or more sections of the AMPD and were rated by more than one person, including an appropriate statistical measure of inter-rater reliability (IRR). The appropriate statistical measure of IRR was judged to be the single-rater ICC or an equivalent statistic. (b) Studies were required to have taken place in clinical or non-clinical settings (no exclusion criteria was applied for setting), (c) published in peer-reviewed journals only, and (d) written or translated into the English Language

*Exclusion Criteria*

 Studies were excluded in the following specific circumstances: (a) based on design (qualitative studies, theoretical, conceptual, or critical commentary, reviews or meta-analyses); (b) population (under 18s not included); or(c) outcome (outcomes not conceptually relevant to IRR or insufficient reporting of statistical data).

**Identification and Selection of Studies**

Articles were identified, screened, and assessed using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework (Moher et al., 2009) (Figure 1). Papers were extracted from electronic databases using reference management software and duplicate articles were removed (Lorenzetti and Ghali, 2013).

Titles and abstracts of the remaining papers were screened by the primary author according to eligibility for the review. Ten percent of these papers were independently reviewed by a second rater. The kappa value at title and abstract screening was κ =0.83 (95% CI, [.65 to 1.01], *p* < .001) indicating a strong level of agreement between raters (McHugh, 2012). Any disagreements were discussed with the researcher’s academic supervisor and a final decision on the papers was agreed upon.

Full articles of remaining papers were acquired and assessed for eligibility. The first author and a Trainee Clinical Psychologist read 14 randomly selected papers (ten percent of identified articles) and independently completed an screening checklist to ensure accurate selection of the eligible papers (Boland et al., 2017). The kappa value at full article screening was κ =0.85 (95% CI, [.57 to 1.13], *p* < .001) indicating a strong level of agreement between raters (McHugh, 2012). Discrepancies were discussed and resolved, and whilst a third reviewer was available for consultation, this was not required. The final studies were checked against eligibility criteria to reduce bias. Both raters agreed that all the selected studies met the eligibility criteria.

**Data Extraction and Coding**

A data extraction database was used to record the following information for accepted studies; (a) article details (for example, author, publication year, title, journal), (b) study design and setting (c) sample description (including sample size and demographics), (d) rater description (including number, profession, and training) (e) AMPD criterion assessed, and instruments used, and (f) IRR statistics type and outcome.

 Originally, only single-rater ICCs were extracted from studies as this was the most commonly reported and clinically relevant statistically. A decision was later taken to also include weighted kappa in the analysis due to statistical similarity with single-rater ICC (Schuster, 2004). Where only “mean of *k* raters” ICCs were reported, studies were excluded from final analysis. Data duplication was monitored throughout, and a further two studies were excluded as a result.

**Quality Assessment and Risk of Bias**

The quality and risk of bias of included studies was assessed using the Quality Appraisal of Reliability Studies (QAREL; Lucas et al., 2010). We note that in the original pre-registered design, a different tool had been proposed, but the QAREL was subsequently identified as an instrument that more directly addressed the primary questions in the study. The QAREL is an 11-item checklist used to assess reliability studies. It explores seven principles representing the appropriateness of subjects, qualification of examiners, examiner blinding, order effects of examination, suitability of the time interval between repeated measurements, appropriate test application and interpretation, and statistical analysis. The scoring criteria for each item in each of the domains are Yes, No, Unclear, or N/A. Each item on the checklist is weighted equally. The total number of “yes” values are calculated as a percentage. A second reviewer was consulted where a judgement was unclear. High quality studies were defined as ≥60% “yes” on the QAREL (Cuchna et al., 2016).

**Data synthesis**

As noted, given the limited available data in relation to any elements of the AMPD beyond Criterion A (only three studies reported information for Criterion B, and three studies reported information in relation to overall diagnosis), there was judged to be no meaningful way to synthesize these elements and therefore the study focused on the synthesis of IRR values in relation to Criterion A only by means of a meta-analytic approach.

Meta-analyses were conducted using the Metafor package version 3.0-2 (Viechtbauer, 2010) in R Studio version 1.4.1717 (Wallace at al., 2012). Random effects models were used due to the presumed variance in effect sizes extracted from each study (Cuijpers, 2016). This approach allows for consideration of differences in true effect sizes between studies, as it provides broader and more conservative 95% confidence intervals than fixed effects models.

For each meta-analysis, ICC and sample size values were extracted from each study, transformed to Fisher's *z* scale, and combined using a random–effects models (Cuchna et al., 2016; Borenstein et al., 2009). Fisher's *z* transformations are important to account for the non- normal distribution in these types of statistics (Cuijpers, 2016). Heterogeneity within meta- analyses was assessed following the transformation to Fisher's *z* by inspecting forest plots as well as using Cochran’s Q test (Cochran, 1954) and the *I*2 statistic (Higgins & Thompson, 2002). The Q test is important to determine whether heterogeneity was significant and the *I*2 statistic provides a percentage of variation across studies due to heterogeneity versus chance. To aid in the interpretation of the results, Fisher's *z* values were then converted back to ICC values after completing meta-analyses (Cuchna et al., 2016).

**Sub-group & Sensitivity Analysis**

Sub-group analysis for Total LPFS Score was conducted using random effects model for instrument type (instrument designed for use with the AMPD vs no AMPD instrument).

 Sensitivity analysis was conducted to examine whether results were skewed by studies with a high risk of bias. High risk studies were removed if this was judged to be the case. Publication bias was explored visually using a funnel plot and the “trim and fill” method was applied to estimate IRR after bias had been accounted for (Duval &Tweedie, 2000; Viechtbauer, 2010).

**Results**

The selection process is outlined in the PRISMA diagram (Figure 1). The search yielded 719 studies of which 337 were excluded as duplicates resulting in 386 articles which were screened based on title and abstract. During this stage, 134 articles were subject to full eligibility screening, resulting in a total of 15 eligible studies. The most common reason for exclusion at the final stage was due to an AMPD criterion not being rated by two or more raters. 14 studies reported IRR statistics for LPFS total score and 11 reported statistics for the individual domains. Total sample size across LPFS Total Score was 902 (range from six to 162), this reduced to 775 for the individual domains (range 10 to 162).

**Study Characteristics**

Characteristics (location, instrument, methods, sample size, sample demographic [age and gender], population type and rater information [type and number]) for the 15 studies included in the review can be found in Table 1. Rater experience varied across studies from clinically inexperienced lay raters (Zimmerman et al., 2014) to qualified mental health professionals (Morey, 2019). Ethnicity of sample and raters was largely unreported across studies, therefore is not included in Table 1.

**Study Design**

Studies used a variety of methods to assess IRR. Three studies used written vignettes or accounts of life stories. The remaining studies utilized a mix of live interviews and video recordings. Instruments used for assessments varied considerably. Three studies used Structured Clinical Interviews (SCIs) designed for assessing the categorical system of PD diagnosis (Structured Interview of Personality Organization; STIPO [Clarkin et al., 2004]) and the SCID–II (First et al., 1995) and eight used a specific AMPD assessment tool (SCID–5 AMPD; Bender et al., 2018), The Clinical Assessment of the Level of Personality Functioning Scale (CALF; Thylstrup et al., 2016) and the Interview for Personality Functioning DSM–5 (STiP-5.1; Hutsebaut et al., 2014). Other studies did not use an instrument, but instead relied on clinical interview (or reviewing materials) using the information provided in the AMPD to guide assessments.

**Risk of Bias Assessment**

Outcome of the QAREL (Lucas et al., 2010) can be seen in Table 1. All studies bar one reached the 60% cut off for low risk of bias and therefore high quality. Roche et al. (2018) scored 54.5% and was therefore rated as high risk of bias and low quality. This study was included in the initial analysis, however removed in sensitivity analysis.

**Criterion A – LPFS**

A total of 14 studies reported 17 interrater reliability scores using single-rater ICCs or equivalent for Total LPFS score (one study reported three IRR tests using independent pools of raters, and one study reported two separate analyses based on two separate instruments conducted in separate interviews, with separate samples). This resulted in a pooled ICC of .75 (95% CI .63 – .84), however this was significantly heterogeneous (*Q* (16) =171.18, *p*<.01, *I*2= 90.10%; see Figure 2). This is above the DSM-5 cut off for acceptable IRR for the DSM and would be categorized as good reliability under ICC reporting guidelines (Kraemer et al., 2012; Koo & Li, 2016). Table 3 provides additional information on overall agreement.

# Subgroup analysis of Total LPFS Score

Subgroup analysis of the assessment instrument can be found in Table 3. This separated studies into those which used an instrument designed for use with the AMPD (i.e., SCID-5-AMPD, StiP-5.1 and CALF) and those that either used an alternative structured assessment not recommended for the AMPD, or no assessment tool at all. Results showed the AMPD instruments yielded higher IRR correlations (*ICC*=.87, 95% CI [.79, .92]; Figure 3), however heterogeneity remained high for this group. Conversely, in the second group (*ICC*=.51, 95% CI [.42, .60]; Figure 4), heterogeneity was no longer significant (*Q* (8) =9.98, *p*=0.27, *I*2 = 25.30%; see Table 2).

# Sensitivity analysis of Total LPFS estimates

Sensitivity analysis was conducted by removing studies rated as having a high risk of bias (n=1) (Roche et al., 2018). This increased reliability ICC to .80 (95% CI .70 – .87) and decreased heterogeneity somewhat, although this remained overall high. Leave-one-out analysis (Viechtbauer, 2010) was also performed which showed no single study accounted for the high levels of heterogeneity.

**Publication bias of Total LPFS estimates**

Publication bias was assessed using funnel plots of Total LPFS IRR. Funnel plots appeared symmetrical and trim and fill analysis did not highlight any missing studies. Visual inspection of forest plots (Figure 2) highlighted small sample studies accounted for large confidence intervals (Kampe et al., 2018; Preti et al., 2018; Morey, 2019; Zimmerman et al., 2014).

**LPFS Domains**

Eleven studies reported single-rater ICCs for the individual LPFS domains (again, one study reported three IRR tests using independent pools of raters). The main findings for each individual domain meta-analysis can be found in Table 3. This table provides information on the number of studies (k), pooled sample size (N), estimate of overall ICC (ICC), 95% confidence intervals, significance test of weighted effect size estimate (z) and amount of heterogeneity (Q). Significant heterogeneity was found in all domains with pooled ICCs ranging from empathy (ICC= .63, 95% CI [.43 – .77]), to intimacy (ICC= .73, 95% CI [.54 –.85]). All pooled domains were above the DSM-5 cut off for acceptability and fell within the moderate category of reliability according to ICC reporting guidelines (Kraemer et al., 2012; Koo & Li, 2016).

**Sensitivity analysis of LPFS domain estimates**

Sensitivity analysis was conducted by removing domain studies rated as having high risk of bias (i.e., Roche et al., 2018). This increased reliability (ICCs ranged from .73, 95% CI [.57-.84] to .82, 95% CI [.68-.90]), however heterogeneity remained significant. Leave-one-out analysis (Viechtbauer, 2010) was performed and showed no one study accounted for the high levels of heterogeneity.

.

**Publication bias of LPFS Domain estimates**

Publication bias was assessed using funnel plots of all domains. Funnel plots appeared symmetrical, however trim and fill analysis highlighted one missing study for the *Identity* domain. However, pooled ICC for this domain remained at .70, 95% CI [0.54, 0.81] even when this study was accounted for. Visual inspection of forest plots (Figures 3-6) showed small sample studies accounted for large confidence intervals (Preti et al., 2018; Zimmerman et al., 2014).

# Discussion

**Summary of Findings**

This review provides a summary of the literature concerning the IRR of the LPFS and hence Criterion A of the AMPD. Fifteen studies were included in the current review. ICCs in Criterion A and its individual domains were meta-analysed to provide a pooled overall ICC. Sub-group and sensitivity analyses were carried out to explore high levels of heterogeneity and possible causes.

Meta-analysis found overall “good” IRR for the LPFS score across the total sample (*ICC*=.75; 95% CI [.63 – .84]). This is above the acceptable threshold for inclusion in the DSM-5 (Kraemer et al., 2012) and higher than previous estimates of IRR of the categorical model (*k*=.40 to .52; Samuel, 2015). However, there was significant heterogeneity across these studies (*I*2 =90.10%). Subgroup analysis highlighted considerable difference between assessment methods; assessment tools designed for the AMPD produced substantially higher IRR scores (*ICC=.87,* 95% CI [.79, .92]) than those that did not (*ICC*=.51, 95% CI [.42, .60]). Whilst heterogeneity was high, this potentially adds weight to the argument that reliability is improved with the use of an SCI (Aboraya et al., 2006; Wood et al., 2002). This may also indicate inconsistency in either application or understanding of the model when relying primarily on clinical judgement, or when using a non-specific SCI (Monahan, 1981; Kitamura & Kitamura, 2000).

Whilst the findings do generally seem to therefore provide broad support for the use of the LPFS, it must be noted that there was also substantial methodological heterogeneity between studies that arguably impinges on the extent to which these findings can be generalized. These differences extend to the method used to present the clinical material, the extent to which the clinical material represented the clinical reality of assessing people presenting with personality concerns, and the qualifications, experience and heterogeneity of the sample of raters utilized. Depending on the aims of the study, a methodologically stronger approach in one element was often associated with a ‘trade off’ in another. For instance, some studies seemed to focus on providing a broader range of *raters* whilst limiting the nature of the clinical material presented (e.g. Morey et al 2019 used a large sample of 123 mental health professionals to rate written vignettes), whereas other studies focused on providing a broader range of clinical material, with a correspondingly smaller number of raters being involved (e.g. Roche et al 2019 used information from 240 students, rated by a smaller panel of trained Research Assistants).

Indeed, in regard to the method used to present clinical material, there were broad differences observed between studies. Three studies used written vignettes, and two audio recordings, which are arguably relatively poor approximations of the clinical information that clinicians would process during a real-life clinical assessment. On the other hand, other studies used video recordings, or in some cases live interviews, and it was common for a mixture of approaches to be adopted.

In regard to the qualifications and experience of the raters, there were significant differences between studies. Some studies utilized professional clinicians (e.g. psychologists or psychiatrists), some utilized unqualified but trained raters (e.g. Research Assistants), and others utilized untrained students or community members. In the latter case, this was typically with the aim of demonstrating the ability of the structured tool to be adopted by a wide audience; however, of course, the extent to which these findings are convincing or relevant in clinical terms is impacted to a large degree by the extent to which the clinical material is representative of the information that would be available to a rater in routine clinical practice; demonstrating interrater reliability presumably becomes much easier with much more contrived and standardized information. In this regard, it is noted there were unfortunately many differences also. Indeed, even when the material being rated was based on patients within clinical services, there was variation in the degree to which one might have reasonably believed the population to have warranted consideration for a personality disorder. Arguably the best approaches (e.g. Hutsebaut et al, 2017) utilized clinical samples where assessment of personality disorder would have likely happened in any case. Other studies used clinical samples where this was plausible but less clear (for instance, in the ‘clinical sample’ recruited by Zettl et al 2019 the most predominant diagnosis was ‘mood disorder’ and only 51% of the sample had been identified as having a personality disorder prior to the study). Finally, some studies utilized ‘non-clinical’ samples either in combination with clinical samples, or used samples entirely outside a clinical context (for instance Roche et al, 2018, who used a large sample of students). Finally, a concerning finding was the lack of reporting of key demographic variables such as ethnicity, arguably a key feature that could be highly relevant to future research given the lack of invariance demonstrated in some self-reported PD measures (Bagby et al, 2022).

In all, these methodological considerations may be important for future reviews addressing a similar topic. In this regard, one practical recommendation would be that any subsequent review of a process of diagnostic assessment of personality disorders would not only pay attention to quality appraisal in relation to the risk of bias using a more generic instrument such as the QAREL, but also, to a much greater degree, in the extent to which a study’s methodological choices about the sample being rated, the process of conducting the ratings, and the qualifications and practices of the evaluators, matching routine clinical practice. This could be conducted either through an adapted quality appraisal process or, at least to some extent, through additional subgroup analyses as further studies become available.

**LPFS Individual Domains**

One potentially intriguing finding highlighted within the study relates to the variability in pooled agreement across the domains, with *empathy* reporting the lowest pooled agreement and *intimacy* the highest (Table 3). Whilst acknowledging the high heterogeneity across domains, this suggests some domains may be more reliably agreed upon than others. Several factors could be relevant in understanding these apparent differences. Firstly, there may be practical difficulties in assessing the domains. For example, intimacy dysfunction may be assessed through self or observer reports of relationship breakdowns. On the other hand, assessment of empathy may be more challenging, arguably requiring specific attention towards, and perhaps skill in assessing, reflective functioning (Zimmerman et al., 2014). It is also possible that the impact of relational difficulties is more salient to the person being assessed, and therefore perhaps more obviously or naturally highlighted in response to questions. Finally, the variation in IRR could highlight fundamental differences in the way the domains are understood by raters. This could be attributed to lack of familiarity with the concepts, difficulties in defining the concepts (particularly across cultures) or indeed flaws within the concepts themselves. This aligns with previous critiques of the model suggesting the domains are too complex to be universally understood (Pincus, 2011; Pilkonis et al., 2011). Taken together, these factors may highlight a need for additional training and or the use of a specific AMPD SCI to guide this process.

**Limitations**

To the best of the authors’ knowledge, this is the first meta-analytic review of the IRR of the AMPD. However, this review does present with several limitations, in addition to the methodological challenges within the studies already considered. It is quite likely that these methodological differences would go a long way to explain the high levels of heterogeneity within all of the meta-analyses conducted.

Secondly, methodological limitations within the review itself are also important to consider. Searches were limited to English-language articles with translations not available despite contacting authors. Only published peer-review articles were included, which meant unpublished or grey literature was potentially excluded. Furthermore, data extraction, coding and quality rating were mainly conducted by a single author, which could have led to reporting bias.

Thirdly, it is important to consider limitations around reporting of IRR statistics within studies. For the dimensional criteria, single-rater ICCs were the most frequent and appropriate reporting method. However, some studies were excluded for reporting different forms of ICC or other IRR statistics, with raw data not available. This was an important methodological decision to correctly pool data due to ICC of the “mean of *k* raters” always appearing larger than the corresponding single-rater type (Koo & Li, 2016). This was also more clinically appropriate as diagnoses in clinically practice are unlikely to be independently rated by a team of clinicians, so this inflated IRR is unlikely to be replicated in clinical practice.

Fourth, trim and fill analysis highlighted a possible missing study within the *Identity* domain which could indicate publication bias (Duval & Tweedie, 2000; Viechtbauer, 2010). However, correcting for this had no impact on the overall ICC score. No missing studies were found on any other domains. Removal of studies with high risk of bias improved domain scores (*Empathy ICC*= .73, 95% CI [.57-.84]; to *Intimacy ICC*=.82, 95% CI [.68-.90]), however variability between scores remained and heterogeneity remained significant.

Finally, but perhaps most substantially, we acknowledge that the review was only able to consider reliability in relation to Criterion A, with insufficient studies available to conduct any meaningful analysis of actual trait domains or overall diagnosis. As new research is published, there is certainly scope for a future review to consider such an analysis, which would also ideally give attention to the methodological concerns raised presently.

**Clinical Implications**

The findings of this study, whilst limited, are clinically important. Results indicate the pooled ICCs of the LPFS and individual domains all fall above the acceptable level for inclusion in DSM-5 (Kraemer et al., 2012). Furthermore, pooled ICCs suggested “good” reliability for LPFS and “moderate” reliability of the individual domains. It is hoped this will go some way to support the case for inclusion into future DSM editions. Furthermore, sub-group analysis shows that using a SCI designed for the AMPD (at least for Criterion A) improves reliability even further. This supports previous research and practice guidelines which recommend the use of SCIs in clinical practice (Samuel, 2015; Wood et al., 2002). It also advises caution around the use of unstructured clinical interviews. This mirrors previous research which suggests clinicians overestimate their ability in other areas of clinical decision making (Monahan, 1981; Kitamura & Kitamura, 2000).

The results also show promising data for the severity domain of the AMPD and its learnability for clinicians, particularly considering some studies opted for inexperienced or untrained raters. However, it is important to highlight that even in research conditions, few “excellent” IRRs were reported across any domain. Taken together, this information may highlight the need for more consistent use of SCIs within clinical practice, additional training, or clarification around AMPD constructs and indicate the need for further research.

**Future Research**

Future research of IRR in Criterion B and overall diagnosis is clearly urgently warranted. Given tentative findings of this review, it would be beneficial to focus research using AMPD SCIs. For overall diagnosis, a particular focus on other presentations aside from Borderline PD would be beneficial. Furthermore, although the inclusion of inexperienced raters in research has been helpful to highlight learnability of the AMPD, further research focusing on implementation of the model in clinical practice would be advantageous to increase ecological findings. Similarly, as already highlighted, future reviews should consider the variation in methodology adopted in existing studies, and reporting key information such as ethnicity. Future reviews would also be assisted if subsequent IRR studies reported both single-rater and “mean of *k* raters” ICCs for more accurate comparisons between studies to be made. Finally, it may be helpful to conduct this review again focusing on test-retest reliability which is considered another fundamental element of reliability alongside IRR (Kraemer et al., 2012).

# References

Aboraya, A., Rankin, E., France, C., El-Missiry, A., & John, C. (2006). The Reliability of Psychiatric Diagnosis Revisited: The Clinician’s Guide to Improve the Reliability of Psychiatric Diagnosis. *Psychiatry*, *3*(1), 41-50.

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders: DSM-5*. (5th ed.) Arlington (VA): American Psychiatric Association.

Bagby, R. M., Keeley, J. W., Williams, C. C., Mortezaei, A., Ryder, A. G., & Sellbom, M. (2022). Evaluating the measurement invariance of the Personality Inventory for DSM-5 (PID-5) in Black Americans and White Americans. Psychological Assessment, 34(1), 82-90. <https://doi.org/10.1037/pas0001085>

Bender, D. S., Morey, L. C., & Skodol, A. E. (2011). Toward a Model for Assessing Level of Personality Functioning in DSM–5, Part I: A Review of Theory and Methods. *Journal of Personality Assessment*, *93*(4), 332–346. <https://doi.org/10.1080/00223891.2011.583808>

Bender, D. S., Skodol, A. E., First, M. B., & Oldham, J. M. (2018). Module I: Structured Clinical Interview for the Level of Personality Functioning Scale. In M. B. First, A. E. Skodol, D. S. Bender, & J. M. Oldham (Eds.), Structured Clinical Interview for the DSM-5 Alternative Model for Personality Disorders (SCID-5-AMPD). *American Psychiatric Association.*

Boland, A., Cherry, G., & Dickson, R. (Eds.). (2017). *Doing a systematic review: A student's guide.* London: Sage.

Borenstein, M., Hedges, L. v., Higgins, J. P. T., & Rothstein, H. R. (2009). Introduction to meta-analysis. *Introduction to Meta-Analysis*, 1–421. <https://doi.org/10.1002/9780470743386>

Buer Christensen, T., Paap, M. C. S., Arnesen, M., Koritzinsky, K., Nysaeter, T.-E., Eikenaes, I., Germans Selvik, S., Walther, K., Torgersen, S., Bender, D. S., Skodol, A. E., Kvarstein, E., Pedersen, G., & Hummelen, B. (2018). Interrater reliability of the Structured Clinical Interview for the DSM–5 Alternative Model of Personality Disorders Module I: Level of Personality Functioning Scale. *Journal of Personality Assessment*, *100*(6), 630–641. <https://doi.org/10.1080/00223891.2018.1483377>

Caldwell-Harris, C. L., & Ayçiçegi, A. (2006). When Personality and Culture Clash: The Psychological Distress of Allocentrics in an Individualist Culture and Idiocentrics in a Collectivist Culture. *Collectivism (Particularly Vertical),* 43(3), 331–361. https://doi.org/10.1177/1363461506066982

Chinese Psychiatry Association. (2000). *Chinese Classification of Mental Disorders. 3rd ed*. Shandong Science Press, Jinan.

Clarkin, J. F., Caligor, E., Stern, B., & Kernberg, O. F. (2004). Structured interview of personality organization (STIPO). Personality Disorders Institute. *Weill Medical College of Cornell University.*

Clarkin, J. F., & Huprich, S. K. (2011). Do DSM-5 Personality Disorder Proposals Meet Criteria for Clinical Utility? *Journal of Personality Disorders. 25*(2), 192–205. https://doi.org/10.1521/PEDI.2011.25.2.192

Cochran, W. (1954). The Combination of Estimates from Different Experiments. Biometrics, 10(1), 101-129. https://doi.org/10.2307/3001666

Costa, P. T., Jr., & Widiger, T. A. (Eds.). (1994). *Personality disorders and the five-factor model of personality.* American Psychological Association. https://doi.org/10.1037/10140-000

Crawford, M. J., Koldobsky, N., Mulder, R., & Tyrer, P. (2011). Classifying Personality Disorder According to Severity. *Journal of Personality Disorders. 25*(3), 321–330. <https://doi.org/10.1521/PEDI.2011.25.3.321>

Cuchna, J. W., Hoch, M. C., & Hoch, J. M. (2016). The interrater and intrarater reliability of the functional movement screen: A systematic review with meta-analysis. *Physical Therapy in Sport*, *19*, 57–65. https://doi.org/10.1016/J.PTSP.2015.12.002

Cuijpers, P. (2016). Meta-analyses in mental health research: A practical guide. Amsterdam, Netherlands, Arije universiteit Amsterdam.

Cruitt, P. J., Boudreaux, M. J., King, H. R., Oltmanns, J. R., & Oltmanns, T. F. (2019). Examining Criterion A: DSM–5 level of personality functioning as assessed through life story interviews. *Personality Disorders: Theory, Research, and Treatment*, *10*(3), 224–234. https://doi.org/10.1037/per0000321

Dereboy, F., Dereboy, C., & Eskin, M. (2018). Validation of the DSM-5 Alternative Model Personality Disorder Diagnoses in Turkey, Part 1: LEAD Validity and Reliability of the Personality Functioning Ratings. *Journal of Personality Assessment*, *100*(6), 603– 611. <https://doi.org/10.1080/00223891.2018.1423989>

Duval, S., & Tweedie R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*. 56(2):455-463. <https://doi.org/10.1111/j.0006-341x.2000.00455.x>

Eysenck, H. J., & Eysenck, S. B. G. (1975). *Manual of the Eysenck Personality Questionnaire (Junior and Adult)*. London: Hodder and Stoughton.

First, M. B., Spitzer, R. L., Gibbon, M., Williams, J. B. W., Davies, M., Borus, J., Howes, M. J., Kane, J., Pope, H. G., & Rounsaville, B. (2011). The Structured Clinical Interview for DSM-III-R Personality Disorders (SCID-II). Part II: Multi-Site Test-Retest Reliability Study. *Journal of Personality Disorders*, *9*(2), 92–104. https://doi.org/10.1521/PEDI.1995.9.2.92

Garcia, D. J., Skadberg, R. M., Schmidt, M., Bierma, S., Shorter, R. L., & Waugh, M. H. (2018). It’s not that Difficult: An Interrater Reliability Study of the DSM-5 Section III Alternative Model for Personality Disorders. *Journal of Personality Assessment*, *100*(6), 612–620. https://doi.org/10.1080/00223891.2018.1428982

Hemmati, A., Rahmani, F., & Bach, B. (2021). The ICD-11 Personality Disorder Trait Model Fits the Kurdish Population Better Than the DSM-5 Trait Model. *Front. Psychiatry* 12:635813. <https://doi.org/10.3389/fpsyt.2021.635813>

Higgins, J. P. T. (2008). Commentary: Heterogeneity in meta-analysis should be expected and appropriately quantified. *International Journal of Epidemiology*, 37(5), 1158– 1160. <https://doi.org/10.1093/ije/dyn204>.

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, 21(11), 1539–1558. <https://doi.org/10.1002/sim.1186>

Hutsebaut, J., Berghuis, H., De Saeger, H., Kaasenbrood, A., & Ingenhoven, T. (2014). Semi-structured Interview for Personality Functioning DSM-5 (STiP-51). The Podium DSM-5 Research Group of the Netherlands Centre of Expertise on Personality Disorders. Utrecht, the Netherlands: Trimbos Institute.

Hutsebaut, J., Kamphuis, J. H., Feenstra, D. J., Weekers, L. C., & de Saeger, H. (2017). Assessing DSM5-oriented level of personality functioning: Development and psychometric evaluation of the semi-structured interview for personality functioning DSM5 (STiP-5.1). *Personality Disorders: Theory, Research, and Treatment*, 8(1), 94–101. <https://doi.org/10.1037/per0000197>

Hutsebaut, J., Weekers, L. C., Tuin, N., Apeldoorn, J. S. P., & Bulten, E. (2021). Assessment of ICD-11 Personality Disorder Severity in Forensic Patients Using the Semi-structured Interview for Personality Functioning DSM-5 (STiP-5.1): Preliminary Findings. *Frontiers in Psychiatry*, *12*. <https://doi.org/10.3389/fpsyt.2021.617702>

Kampe, L., Zimmerman, J., Bender, D., Caligor, E., Borowski, A.-L., Ehrenthal, J. C., Benecke, C., & Hörz-Sagstetter, S. (2018). Comparison of the structured DSM–5 clinical interview for the Level of Personality Functioning Scale with the Structured Interview of Personality Organization. *Journal of Personality Assessment*, *100*(6), 642–649. <https://doi.org/10.1080/00223891.2018.1489257>

Kitamura, T., & Kitamura, F. (2000). Reliability of clinical judgment of patients’ competency to give informed consent: A case vignette study. *Psychiatry and Clinical Neurosciences*, *54*(2), 245–247. https://doi.org/10.1046/J.1440-1819.2000.00665.X

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, *15*(2), 155. https://doi.org/10.1016/J.JCM.2016.02.012

Kraemer, H. C., Kupfer, D. J., Clarke, D. E., Narrow, W. E., & Regier, D. (2012). DSM-5: How Reliable Is Reliable Enough? *The American Journal of Psychiatry. 169*(1), 13– 15. https://doi.org/10.1176/APPI.AJP.2011.11010050

Krueger, R. F., Skodol, A. E., Livesley, W. J., Shrout, P. E., & Huang, Y. (2007). Synthesizing dimensional and categorical approaches to personality disorders: refining the research agenda for DSM‐V Axis II. *International Journal of Methods in Psychiatric Research*, *16*(1), S65. <https://doi.org/10.1002/MPR.212>

Lorenzetti, D.L., Ghali, W.A. (2013) Reference management software for systematic reviews and meta-analyses: an exploration of usage and usability. *BMC Med Res Methodol* **13,** 141 <https://doi.org/10.1186/1471-2288-13-141>

Lucas, N. P., Macaskill, P., Irwig, L., & Bogduk, N. (2010). The development of a quality appraisal tool for studies of diagnostic reliability (QAREL). *Journal of Clinical Epidemiology*, *63*(8), 854-861. <https://doi.org/10.1016/j.jclinepi.2009.10.002>

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, *22*(3), 276–282. <https://doi.org/10.11613/BM.2012.031>

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ*, *339*(7716), 332– 336. https://doi.org/10.1136/BMJ.B2535

Monahan, J. (1981). *Predicting Violent Behavior - An Assessment of Clinical Techniques.* Office of Justice Programs, Sage, Beverly Hills, CA.

Morey, L. C. (2019). Interdiagnostician Reliability of the DSM-5 Section II and Section III Alternative Model Criteria for Borderline Personality Disorder. *Journal of Personality Disorders*, *33*(6), 721–863. https://doi.org/10.1521/pedi\_2019\_33\_362

Mulder, R. T. (2021). ICD-11 Personality Disorders: Utility and Implications of the New Model. *Frontiers in Psychiatry*, *12*, 709. <https://doi.org/10.3389/fpsyt.2021.655548>

Mulder, R., & Tyrer, P. (2019). Diagnosis and classification of personality disorders: novel approaches. *Current Opinion in Psychiatry*, *32*(1), 27-31. <https://doi.org/10.1097/yco.0000000000000461>

Pilkonis, P. A., Hallquist, M. N., Morse, J. Q., & Stepp, S. D. (2011). Striking the (Im)proper balance between scientific advances and clinical utility: Commentary on the DSM-5 proposal for personality disorders. *Personality Disorders: Theory, Research, and Treatment*, *2*(1), 68–82. https://doi.org/10.1037/A0022226

Pincus, A. L. (2011). Some comments on nomology, diagnostic process, and narcissistic personality disorder in the DSM-5 proposal for personality and personality disorders. *Personality Disorders: Theory, Research, and Treatment*, *2*(1), 41–53. https://doi.org/10.1037/A0021191

Roche, M. J., Jacobson, N. C., & Phillips, J. J. (2018). Expanding the Validity of the Level of Personality Functioning Scale Observer Report and Self-Report Versions Across Psychodynamic and Interpersonal Paradigms. *Journal of Personality Assessment,* 100(6), 571-580. https://doi.org/10.1080/00223891.2018.1475394

Preti, E., di Pierro, R., Costantini, G., Benzi, I. M. A., de Panfilis, C., & Madeddu, F. (2018). Using the Structured Interview of Personality Organization for DSM–5 Level of Personality Functioning Rating Performed by Inexperienced Raters. *Journal of Personality Assessment*, *100*(6), 621–629. https://doi.org/10.1080/00223891.2018.1448985

Samuel, D. B. (2015). A review of the agreement between clinicians’ personality disorder diagnoses and those from other methods and sources. *Clinical Psychology: Science and Practice,* 22(1), 1–19. https://doi.org/10.1111/cpsp.12088

Sansone, R. A., & Sansone, L. A. (2011). Gender Patterns in Borderline Personality Disorder. *Innovations in Clinical Neuroscience*, *8*(5), 16-20.

Schuster, C. (2004). A Note on the Interpretation of Weighted Kappa and its Relations to Other Rater Agreement Statistics for Metric Scales. *Educational and Psychological Measurement*, *64*(2), 243–253. https://doi.org/10.1177/0013164403260197

Somma, A., Borroni, S., Carlotta, D., Giarolli, L. E., Barranca, M., Cerioli, C., Franzoni, C., Masci, E., Manini, R., Busso, S. L., Ruotolo, G., & Fossati, A. (2019). The inter-rater reliability and convergent validity of the Italian translation of the Structured Clinical Interview for the DSM-5 Alternative Model of Personality Disorders Module III in a psychotherapy outpatient sample. *Journal of Psychopathology*, *25*(4), 195–204.

Somma, A., Borroni, S., Gialdi, G., Carlotta, D., Emanuela Giarolli, L., Barranca, M., Cerioli, C., Franzoni, C., Masci, E., Manini, R., Luca Busso, S., Ruotolo, G., Krueger, R. F., Markon, K. E., & Fossati, A. (2020). The Inter-Rater Reliability and Validity of the Italian Translation of the Structured Clinical Interview for DSM-5 Alternative Model for Personality Disorders Module I and Module II: A Preliminary Report on Consecutively Admitted Psychotherapy Outpatients. *Journal of Personality Disorders*, *34* (Supplement C), 95–123. <https://doi.org/10.1521/pedi_2020_34_511>

Thylstrup, B., Simonsen, S., Nemery, C., Simonsen, E., Noll, J. F., Myatt, M. W., & Hesse, M. (2016). Assessment of personality-related levels of functioning: A pilot study of clinical assessment of the DSM-5 level of personality functioning based on a semi- structured interview. *BMC Psychiatry*, *16*(1), 1–8. <https://doi.org/10.1186/s12888-016-1011-6>

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124–1131. <https://doi.org/10.1126/science.185.4157.1124>

Verheul, R., & Widiger, T. A. (2004). A meta-analysis of the prevalence and usage of the personality disorder not otherwise specified (PDNOS) diagnosis. *Journal of Personality Disorders*, *18*(4), 309–319. https://doi.org/10.1521/PEDI.18.4.309.40350

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36(3), 1-48. https://doi.org/10.18637/jss.v036.i03

Wallace, B. C., Dahabreh, I. J., Trikalinos, T. A., Lau, J., Trow, P., & Schmid, C. H. (2012). Closing the Gap between Methodologists and End-Users: R as a Computational Back- End. Journal of Statistical Software, 49(5), 1–15. https://doi.org/10.18637/jss.v049.i05

Westen, D., & Shedler, J. (1999). Revising and assessing Axis II, Part 2: Towards an empirically based and clinically useful classification of personality disorders. *American Journal of Psychiatry*, 156, 273-285.

Widiger, T. A., Livesley, W. J., & Clark, L. A. (2009). An Integrative Dimensional Classification of Personality Disorder. *Psychological Assessment*, *21*(3), 243–255. https://doi.org/10.1037/A0016606

Winsper, C., Bilgin, A., Thompson, A., Marwaha, S., Chanen, A. M., Singh, S. P., Wang, A., & Furtado, V. (2019). *The prevalence of personality disorders in the community: a global systematic review and meta-analysis*. https://doi.org/10.1192/bjp.2019.166

Wood, J. M., Garb, H. N., Lilienfeld, S. O., & Nezworski, M. T. (2002). Clinical assessment. *Annual Review of Psychology, 53*(1), 519-543. https://doi.org/10.1146/annurev.psych.53.100901.135136

World Health Organization. (1992). *The ICD-10 classification of mental and behavioural disorders: Clinical descriptions and diagnostic* guidelines. Geneva: World Health Organization.

Wright, A., & Zimmerman, J. (2015). At the nexus of science and practice: Answering basic clinical questions in personality disorder assessment and diagnosis with quantitativemodeling techniques. Personality Disorders: Toward Theoretical and Empirical Integration in Diagnosis and Assessment., 109–144. <https://doi.org/10.1037/14549-006>

Zimmerman, J., Benecke, C., Bender, D. S., Skodol, A. E., Schauenburg, H., Cierpka, M., & Leising, D. (2014). Assessing DSM–5 Level of Personality Functioning From Videotaped Clinical Interviews: A Pilot Study With Untrained and Clinically Inexperienced Students. 96(4), 397–409. <https://doi.org/10.1080/00223891.2013.852563>

Zettl, M., Volkert, J., Vogele, C., Herpertz, S. C., Kubera, K. M., & Taubner, S. (2020). Mentalization and criterion a of the alternative model for personality disorders: Results from a clinical and nonclinical sample. Personality Disorders, 11(3), 191–201. [https://doi.org/https:/10.1037/per0000356](https://doi.org/https%3A/10.1037/per0000356)



Figure 2 - Forest plot of Total LPFS IRR meta-analysis



Figure 3 - Forest plot of Total LPFS subgroup IRR meta-analysis including AMPD instruments only.

Figure 1 - PRISMA Diagram



Figure 6 - Forest plot for LPFS Domain – Self-Direction IRR meta-analysis

Figure 1- Forest plot of Total LPFS subgroup IRR meta-analysis (non AMPD instruments only)



Figure 2 Forest plot for LPFS Domain – Identity IRR meta-analysis



Figure 8. Forest plot for LPFS Domain – Intimacy IRR meta-analysis

Figure 7. Forest plot for LPFS Domain – Empathy IRR meta-analysis

# Table 1.

*Overview of final studies included in review*

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Study | Location | Instrument | Methods | Sample size | %Female | Sample age (years) | Population | Raters | No. of ratings pp | Risk of bias QAREL | Risk of Bias Rating |
|  |  |  |  |  |  |  |  |  |  | % of yes |  |
|  |  |  |  |  |  | *Range* | *M±SD* |  |  |  |  |  |
| Buer | Norway | SCID-5- | Live | 17 | 65 | 19-59 | 31.6 | Clinical and | 7 experienced | 5 | 72.7 | Low |
| Christensen et |  | AMPD | interviews |  |  |  |  | non-clinical | raters |  |  |  |
| al. (2018) |  |  | and video recordings |  |  |  |  |  |  |  |  |  |
| Cruitt et al. | United | LSI and | Video | 162 | 56.2 | 55-64 | NR | Non-clinical | 9 undergraduate | 3 | 72.7 | Low |
| (2019) | States | LPFS | recordings |  |  |  |  |  | students |  |  |  |
|  |  |  | of LSIs |  |  |  |  |  |  |  |  |  |
| Dereboy et al. | Turkey | SCID-II | Live | 120 | 66.67 | 16-63 | 35.7± | Clinical | Clinicians\*, 3 | 2 | 63.6 | Low |
| (2018) |  | and LPFS | interviews |  |  |  | 12.5 | (inpatient | CP students, 3 |  |  |  |
|  |  |  |  |  |  |  |  | and | academics, 1 |  |  |  |
|  |  |  |  |  |  |  |  | outpatient) | psychiatrist and |  |  |  |
|  |  |  |  |  |  |  |  |  | 2 CPs |  |  |  |
| Garcia et al. | United | LPFS and | Written | 15 | NR | NR | NR | Clinical | 13 CP doctorate | 13 | 63.6 | Low |
| (2018) | States | PDLT–C | vignettes |  |  |  |  |  | students |  |  |  |
| Hutsebaut et | Netherlands | STiP-5.1 | Live | 40\*\* | 66.3 | 16-61 | 33.6± | Clinical | 12 Psychologists  | 2 | 90.9 | Low |
| al. (2017) |  |  | interviews |  |  |  | 12 |  |  & 3 study authors  |  |  |  |
|  |  |  | and video |  |  |  |  |  |  |  |  |  |
|  |  |  | recordings | 18 | 4.2 | 18-60 | 39± | Community | (same as above) |  |  |  |
|  |  |  |  |  |  |  | 14.5 |  |  |  |  |  |
| Hutsebaut et al. (2021) | Netherlands | STiP-5.1 | Interview and | 30 | 10.3 | 21-65 | 38.43±11.70 | Clinical (forensic | “Psychologists or trainees” | 2 | 81.8 | Low |
|  |  |  | observer |  |  |  |  | inpatient) |  |  |  |  |

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Kampe et al. (2018) | Germany | SCID-5- AMPDand STIPO | Interviews and video recordings | 30 | 60 | 16-61 | 32.5±9.77 | Clinical | Author and a CP Masters student | 2 | 63.6 | Low |
| Morey (2019) | United | DSM | Written | 12 | NR | NR | NR | Clinical (not | 123 mental | 40 | 72.7 | Low |
|  | States | TEXT | vignettes |  |  |  |  | all PD) | health |  |  |  |
|  |  |  |  |  |  |  |  |  | professionals |  |  |  |
| Preti et al. | Italy | STIPO | Live | 10 | 100 | NR | 36.6± | Clinical | 10 clinically | 10 | 72.7 | Low |
| (2018) |  |  | interviews |  |  |  | 11.26 | (inpatients) | inexperienced |  |  |  |
|  |  |  | and audio |  |  |  |  |  | undergraduate |  |  |  |
|  |  |  | recordings |  |  |  |  |  | students |  |  |  |
| Roche et al. | United | LFPS | Written | 70\*\*\* | 80 | NR | 20.52± | Non-clinical | 15 research | 5 | 54.5 | High |
| (2018) | States |  | abbreviate |  |  |  | 0.98 | (students) | assistants (3 |  |  |  |
|  |  |  | d LSIs. | 85ª | 81 | NR | 19.92± |  | teams of 5) |  |  |  |
|  |  |  |  |  |  |  | 1.52 |  |  |  |  |  |
|  |  |  |  | 85 ª | 81 | NR | 19.92± |  |  |  |  |  |
|  |  |  |  |  |  |  | 1.52 |  |  |  |  |  |
| Somma et al. | Italy | SCID-5- | Interview | 84 | 53.6 | NR | 36.42± | Clinical | 10 Trainee CPs | 2 | 72.7 | Low |
| (2019) |  | AMPD | and |  |  |  | 12.94 |  |  |  |  |  |
|  |  |  | observer |  |  |  |  |  |  |  |  |  |
| Somma et al. (2020) | Italy | SCID-5- AMPD | Interview and | 88 | 54.5 | NR | 36.47±14.04 | Clinical | CPs with 1–3 years of | 2 | 90.9 | Low |
|  |  |  | observer |  |  |  |  |  | experience |  |  |  |
| Thylstrup et | Denmark | CALF | Interview | 30b | 47.22 | 18-56 | 36 | Clinical | 4 psychologists | 3 | 63.6 | Low |
| al. (2016) |  |  | and video |  |  |  |  |  | and 2 MDs |  |  |  |
|  |  |  | recording | 7 | 100 | 24-45 | 34 | Non-clinical | (2 interviews |  |  |  |
|  |  |  |  |  |  |  |  |  | conducted by a |  |  |  |
|  |  |  |  |  |  |  |  |  | psychology |  |  |  |
|  |  |  |  |  |  |  |  |  | student) |  |  |  |

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Zettl et al. (2020) | Germany | StiP-5.1 | Interview and audio recordings | 50c | 57.27 | NR | 31.44±11.74 | Clinical and non-clinical | Trained researchers | 2 | 63.6 | Low |
| Zimmerman et al. (2014) | Germany | OPD and LFPS | Video recordings | 10 | 100 | NR | 30.8± | Clinical (inpatients) | 22 untrained and clinically | 22 | 72.7 | Low |
|  |  |  |  |  |  |  | 9.6 |  | inexperiencedstudents |  |  |  |

*Note*. CP=Clinical Psychologist; SCID-5-AMPD= SCI for the DSM-5 Alternative Model for Personality Disorders; LSI = Life Story Interviews; LPFS = Level of Personality Functioning Scale; SCID- II = SCI for the Diagnostic and Statistical Manual of Mental Disorders Axis II; CP=Clinical Psychologist; PDLT–C = Clinician Rating Personality Disorder Level and Traits; StiP-5.1= Semi-Structured Interview for Personality Functioning DSM–5; STIPO= Structured Interview of Personality Organisation ; CALF= Clinical Assessment of the Level of Personality Functioning Scale; OPD= Operationalized Psychodynamic Diagnosis system; \*refers to intake clinicians as raters in addition to “second rater”. Unclear how many clinicians took part, therefore this is coded as 2 raters to reduce error. \*\*only 40 participants were included in IRR analysis, however demographics are reported for the total sample of 80. \*\*\*IRR sample of 70, however only 50 participants completed the initial demographic survey. aIRR sample of 85, demographics reported for total 110 sample of study. bIRR sample of 30, demographics reported for total sample of 37. cIRR sample of 50, demographics reported for total sample of 110.

# Table 2

*Interrater reliability statistics for Criterion A- Total LPFS scores and individual domains organized alphabetically by measure, then author.*

**IRR (95% confidence intervals)**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Study** | **N** | **Measure** | **IRR Stat** | **Total LPFS** | **Identity** | **Self- Direction** | **Empathy** | **Intimacy** |
| Hutsebaut et | 58 | StiP-5.1 | ICC (1,1) | .71 | .76 | .64 | .79 | .80 |
| al. (2017) |  |  |  |  |  |  |  |  |
| Hutsebaut et al. (2021) | 30 | StiP-5.1 | ICC (1,1) | .81 | .54 | .77 | .69 | .90 |
| Zettl et al. | 110 | StiP-5.1 | ICC (2,1) | .92 | .93 | .92 | .89 | .93 |
| (2020) |  |  |  |  |  |  |  |  |
| Kampe et al. | 6 | StiPO | ICC (2,1) | .78 | - | - | - | - |
| (2018) |  |  |  |  |  |  |  |  |
| Preti et al. | 10 | StiPO | ICC (2,1) | .42 | .39 | .35 | .28 | .42 |
| (2018) |  |  |  | (.21,.72) | (.19,.71) | (.16, .67) | (.11,.60) | (.21,.73) |
| Buer | 17 | SCID-5- | ICC (2,1) | .96 | .94 | .94 | .9 | .89 |
| Christensen |  | AMPD |  | (.92,.9) | (.88, .98) | (.87, .98) | (.80, .96) | (.80, .96) |
| et al. (2018) |  |  |  |  |  |  |  |  |
| Kampe et al. | 30 | SCID-5- | ICC (2,1) | .93 | .89 | .79 | .92 | .95 |
| (2018) |  | AMPD |  | (.87,.90) |  |  |  |  |
| Somma et al. | 84 | SCID-5- | Kw | .87 | - | - | - | - |
| (2019) |  | AMPD |  |  |  |  |  |  |
| Somma et al. | 88 | SCID-5- | ICC (1,1) | .87 | .83 | .87 | .77 | .88 |
| (2020) |  | AMPD |  |  |  |  |  |  |
| Thyltsrup et | 30 | CALF | ICC (3,1) | .69 | .59 | .72 | .42 | .65 |
| al. (2016) |  |  |  | (.47-.83) | (.33-.76) | (.51-.84) | (.11-.65) | (.42-.80) |
| Cruitt et al. | 162 | LSI and | ICC (1,1) | .56 | .57 | .50 | .47 | .37 |
| (2019) |  | DSM |  |  |  |  |  |  |
| Garcia et al (2018) | 15 | DSM | ICC (2,1) | .81 |  |  |  |  |
| Morey | 12 | DSM | ICC (1,1) | .50 | - | - | - | - |
| (2019) |  |  |  |  |  |  |  |  |
| Roche et al. | 70 | DSM | ICC (2,1) | .58 | .49 | .44 | .26 | .29 |
| (2018a) |  |  |  | (.48,.69) | (.38,.61) | (.33,.56) | (.16,.38) | (18,.41) |
| Roche et al. | 85 | DSM | ICC (2,1) | .42 | .41 | .29 | .23 | .31 |
| (2018b) |  |  |  | (.32,.53) | (.31,.52) | (.19,.40) | (.13,.3) | (.21,.42) |
| Roche et al. | 85 | DSM | ICC (2,1) | .36 | .41 | .21 | .14 | .23 |
| (2018c) |  |  |  | (.26,.47) | (.31,.52) | (.12,.32) | (.06,.24) | (.13,.33) |
| Zimmerman |  | OPD and |  | .51 | .41 | .46 | .25 | .63 |
| et al. (2014) | 10 | DSM | ICC (2,1) | (.31,.70) | (.23,.71) | (.27, 75) | (.12,.55) | (.43, 85) |

# Table 3.

*Meta-analysis outcomes for Interrater Reliability of Criterion A LPFS and its domains, including subgroup analysis.*

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Variable | *k* | ICC | 95%Confidence | SE | *p* | T2 | *z* | *Q* | *df* | *p* | I2 |
|  |  |  |  | LL | UL |  |  |  |  |  |  |  |  |
| TOTAL LPFS |  | 17 | 0.75 | 0.63 | 0.84 | 0.12 | <.001 | 0.44 | 8.27 | 171.18 | 16 | <.001 | 90.10 |
| Instrument | AMPD | 8 | 0.87 | 0.79 | 0.92 | 0.13 | <.001 | 0.32 | 10.30 | 35.61 | 7 | <.001 | 83.90 |
| used | Instrument |  |  |  |  |  |  |  |  |  |  |  |  |
|  | No AMPDinstrument | 9 | 0.51 | 0.42 | 0.60 | 0.06 | <.001 | 0.01 | 8.87 | 9.98 | 7 | 0.27 | 25.30 |
| Domains | Identity | 13 | 0.70 | 0.55 | 0.81 | 0.13 | <.001 | 0.19 | 6.55 | 148.17 | 12 | <.001 | 91.56 |
|  | Self-Direction | 13 | 0.68 | 0.51 | 0.80 | 0.14 | <.001 | 0.21 | 6.04 | 174.81 | 12 | <.001 | 92.12 |
|  | Empathy | 13 | 0.63 | 0.43 | 0.77 | 0.15 | <.001 | 0.24 | 5.11 | 165.17 | 12 | <.001 | 92.91 |
|  | Intimacy | 13 | 0.73 | 0.54 | 0.85 | 0.16 | <.001 | 0.31 | 5.68 | 247.33 | 12 | <.001 | 94.46 |