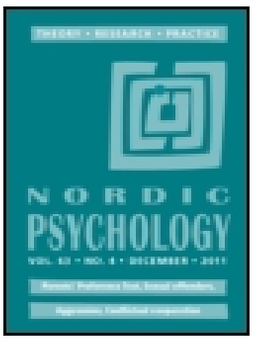

Nordic Psychology

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/rnpy20

Validating measurement tools for mentalization, emotion regulation difficulties and identity diffusion among Finnish adolescents

Sami J. Eloranta , Riittakerttu Kaltiala , Nina Lindberg , Matti Kaivosoja & Kirsi Peltonen

To cite this article: Sami J. Eloranta , Riittakerttu Kaltiala , Nina Lindberg , Matti Kaivosoja & Kirsi Peltonen (2020): Validating measurement tools for mentalization, emotion regulation difficulties and identity diffusion among Finnish adolescents, Nordic Psychology, DOI: 10.1080/19012276.2020.1863852

To link to this article: https://doi.org/10.1080/19012276.2020.1863852

Published online: 30 Dec 2020.

Nordic Psychology, 2020
https://doi.org/10.1080/19012276.2020.1863852

RESEARCH ARTICLE

OPEN ACCESS

Validating measurement tools for mentalization, emotion regulation difficulties and identity diffusion among Finnish adolescents

SAMI J. ELORANTA[1,2], RIITTAKERTTU KALTIALA[1,2], NINA LINDBERG[3,4], MATTI KAIVOSOJA[5,6] & KIRSI PELTONEN[2]

Correspondence address: Sami J. Eloranta, Department of Adolescent Psychiatry, Nuorisopsykiatrian Poliklinikka, Tampere University Hospital, PL 2000, Tampere, 33521, Finland. Email: sami.eloranta@tuni.fi

Abstract

Mentalization, emotion regulation, and identity diffusion are theoretically and clinically important transdiagnostic psychological constructs that contribute to mental health. In order to advance meaningful empirical research on these constructs, we need measures that are well tested. In this study, we used confirmatory factor analysis to assess the reliability and construct validity of the Mentalization Questionnaire (MZQ), different versions of the Difficulties in Emotion Regulation Scale (DERS), and the Assessment of Identity Development and Identity Diffusion in Adolescence (AIDA) with data from a general population of Finnish adolescents (N = 360). For MZQ, the factor structure and validity of the subscales were not confirmed. For DERS, a short version, that did not include the *lack of emotional awareness* subscale was the most coherent and recommendable version of the measure, with a good degree of reliability and a reasonable indication of the convergent and discriminative validity between the subscales. For AIDA, the factor structure was confirmed, but when using this measure for research purposes, it should be taken into account that reverse coding items may affect the factor structure by creating a method factor. The reliability of the AIDA was acceptable, but some of the subscales showed poor convergent and discriminative validity.

Keywords: measure validation, construct validity, mentalization, emotion regulation, identity diffusion, adolescent mental health

Mentalization ability, functional emotion regulation (ER), and coherent identity are all considered core issues in adolescent mental health, and dysfunction in these areas has been suggested as transdiagnostic factors contributing to a wide range of mental health issues

[1]Tampere University Hospital, Tampere, Finland
[2]Tampere University, Tampere, Finland
[3]Helsinki University, Helsinki, Finland
[4]Helsinki University Hospital, Helsinki, Finland
[5]Central Ostrobothnia Central Hospital, Kokkola, Finland
[6]University of Turku, Turku, Finland

(Katznelson, 2014; McLaughlin et al., 2011; Schwartz et al., 2015). In order to effectively study the impact of these theoretical constructs, and to advance the clinical assessment of transdiagnostic factors, there is a need for validated and easily utilizable measures. The aim of this study was to validate the Finnish translations of self-assessment measures for mentalization, difficulties in ER and identity diffusion in a normative adolescent sample and to examine the reliability and construct validity of these measures.

Mentalization, emotion regulation and identity diffusion as transdiagnostic constructs

Mentalization refers to the ability of an individual to understand the inner mental states of themselves and others and to understand the representational nature of the mind (Fonagy & Bateman, 2008). The foundation of this ability to understand the meaning of human behavior that rises from internal states, such as needs, wishes, emotions, beliefs, imagination, and distinguishing inner from outer reality, is believed to develop within an early attachment relationship (Fonagy et al., 2002). Infants seek relationships in order to self-regulate and find meaning. An accurate 'marked mirroring' of affects by a baby's caregiver allows the infant to perceive their own states in the caregiver's mind and to grasp the concept that the caregiver's mind is separate from their own (Fonagy, 2000; Freeman, 2016). Problems in mentalization were originally studied within theory and treatment of Borderline Personality Disorder (BPD), but recently theoretical, clinical, and empirical research on the concept has increased, for example, in relation to psychosis, eating disorders, and parenting (Camoirano, 2017; Fonagy et al., 2011).

ER refers to a set of mental processes that individuals use to dampen, intensify, or maintain emotions, depending on their situational goal (Gross & Thompson, 2007). As with the ability to mentalize, functional ER develops in infancy through interaction with the primary caregiver and the foundation for an individual's adaptive ER strategies is believed to lie in the emotional style and availability of their early attachment figure (Bowlby, 1988; Calkins & Hill, 2007). During adolescence, there seems to be a momentary regression in ER strategies and, subsequently, a reorganization toward adult adaptive ER (Zimmermann & Iwanski, 2014). Dysfunctional ER in adolescence predicts future BPD, anxiety, aggressive behavior, and eating disorders, suggesting that dysfunctional ER poses a risk for a wide range of later psychopathology (Glenn & Klonsky, 2009; McLaughlin et al., 2011).

Identity diffusion is a concept that refers to pathological identity as distinct from a normative identity crisis wherein former identifications and introjections are synthesized into a consolidated identity (Erikson, 1968). Kernberg (2006) claimed that identity diffusion is the key point for differentiating the milder forms of neurotic character pathology from the more severe character pathologies central in personality disorders, particularly BPD. The common elements among the different conceptualizations of identity diffusion are the lack of differentiated and integrated representations of self and others, a negative self-image, a lack of long-term goals, and the lack of a sense of continuity in self-perception over time (Sollberger, 2013).

Measures and their validity

In previous research, mentalization has been measured using the observer-rated reflective functioning (RF) scale (Fonagy et al., 1998). Unfortunately, the RF scale is a complex, time-consuming instrument and is, therefore, difficult to use in everyday clinical practice, or for research purposes with larger datasets. To address this issue, Hausberg et al. (2012) developed a self-reporting questionnaire in order to assess mentalization called the Mentalization Questionnaire (MZQ). The authors noted that using a self-reporting measure for assessing a skill like mentalization is somewhat problematic, but that an individual's attitude toward mentalization might be directly related to their actual ability.

The MZQ has not yet been validated against the more common observer-rated measures, but Hausberg et al. (2012) have stated that patients with suicide attempts, self-injurious behavior, diagnosis of BPD, or multiple diagnosis score lower on MZQ than other patients, and that patients with secure attachment pattern score higher than other patients. They also describe MZQ having a satisfactory internal consistencies (.81 for full scale and .54 to .72 for subscales). A recent study found that reduced mentalization in adolescents measured with MZQ was associated with depression and risk behaviors such as binge drinking, and that the individual's level of mentalization ability mediated the association between childhood traumatic experiences and depression (Belvederi Murri et al., 2017). The Finnish version (Keinänen et al., 2019) of the MZQ was used successfully to assess the changes in mentalization during mentalization-based group therapy among university students with pervasive ER difficulties (Keinänen et al., 2017).

ER has a range of different well-established measures. In particular, the Difficulties in Emotion Regulation Scale (DERS; Gratz & Roemer, 2004) is widely used, has good psychometric qualities, and is reliable and valid in adolescents (Neumann et al., 2010). A Finnish translation (Tapola et al., 2010) has been used with adults, but normative data for adolescents is not available.

The original DERS includes 36 items, with six subscales: *lack of emotional awareness*, *lack of emotional clarity*, *difficulties controlling impulsive behaviors when distressed*, *difficulties engaging in goal-directed behavior when distressed*, *nonacceptance of negative emotional responses*, and *limited access to effective ER strategies*. However, Bardeen et al. (2012) found that the *lack of emotional awareness* subscale does not correlate well with other scales and, thus, suggested a new version of the measure without the subscale (DERS-R). In order to reduce the burden for respondents, the DERS short form (DERS-SF), including half of the original items, was recently introduced. Kaufman et al. (2016) have stated that shorter version shows excellent psychometric properties together with the original structure, and that the factor structure of the measure was more coherent with short form than the original version.

Goth et al. (2012) developed a measure to assess identity diffusion in adolescence called the Assessment of Identity Development and Identity Diffusion in adolescence (AIDA). AIDA is meant to identify the pathological features in adolescent identity that are considered central in personality disorders. The AIDA model differentiates two dimensions of identity: continuity and coherence. The continuity dimension has three subdimensions: *stability in attributes, stability in relations,* and *positive emotional self-reflection*. The coherence dimension also has three subdimensions: *consistent self-image, autonomy,* and *positive cognitive*

self-reflection. Goth et al. (2012) describe very good scale reliabilities (.94 for total scale, .86 and .92 for main dimensions and from .76 to .86 for subdimensions) to AIDA. Jung et al. (2013) have also demonstrated empirically that AIDA can differentiate adolescents with personality disorder from the general population, as well as from adolescents with other types of psychiatric problems. They argue that identity diffusion, as defined in AIDA, is a distinguishing mark of personality disorder, not only psychiatric impairment in general.

When measuring psychological constructs, such as mentalization, ER and identity, the construct validity of the measures should be addressed. Construct validity is seen as an overarching term that includes all other forms of validity. It refers to the extent to which a measure assesses the construct it is supposed to assess (Strauss & Smith, 2009). Campbell and Fiske (1959) introduced the terms *convergent* and *discriminant* construct validity. Convergent validity refers to the degree in which two concepts, that theoretically should be connected, are actually related. Discriminant validity is the degree to which constructs that are theoretically distinct, are in fact, unrelated

Nowadays, an often-used method for examining the construct validity of a measure is confirmatory factory analysis (CFA). CFA allows a comparison of alternative theoretical measurement models at the latent factor level and helps reduce measurement error. It can, therefore, be used to confirm the substructure of the measure (Atkinson et al., 2011; John & Benet-Martínez, 2000; McArdle, 1996).

Aim of the study

This study aimed to evaluate the construct validity and reliability of DERS, MZQ, and AIDA when used with a normative sample of Finnish adolescents. As alternative versions of DERS exist, the study evaluated the construct validity and reliability of the different versions.

Materials and methods

Measures

Mentalization

MZQ (Hausberg et al., 2012) is a 15-item self-reporting measure consisting of four subscales: *refusing self-reflection*, *emotional awareness*, *psychic equivalence mode*, and *regulation of affect*. Other subscales consist of four items, but *regulation of affect* -subscale has three items. For each item, participants choose using five-point likert scale how much they agree or disagree with the item. Hausberg et al. (2012) ascribe the MZQ with satisfactory internal consistencies and good construct validity but recommend using the total score instead of the subscales before further validation. The Finnish translation was conducted by Keinänen et al. (2019).

Emotion regulation

DERS (Gratz & Roemer, 2004) is a self-report measure comprising 36 items and six subscales, labeled as *lack of emotional awareness* (AWARENESS), *lack of emotional clarity* (CLARITY), *difficulties controlling impulsive behaviors when distressed* (IMPULSE), *difficulties engaging in goal-directed behavior when distressed* (GOALS), *nonacceptance of negative emotional*

responses (NONACCEPTANCE), and *limited access to effective ER strategies* (STRATEGIES). Participants choose using five-point likert scale how often the item applies to respondent. Items on one subscale ranges from 5 to 8. The Finnish translation was conducted by Tapola et al. (2010).

Identity diffusion

AIDA (Goth et al., 2012) is a 58-item self-report measure that considers two dimensions of identity: continuity and coherence. Number of items on main dimensions are 27 and 31. The continuity dimension has three subdimensions: *stability in attributes, stability in relations,* and *positive emotional self-reflection*. Coherence has also three subdimensions: *consistent self-image*, *autonomy*, and *positive cognitive self-reflection*. Subdimensions consists of 7 to 12 individual items. For each item, the respondent reports on a five-point scale how well it describes the respondent. The translation of the AIDA was conducted as a part of this study. According to the requirements of the measure's developers it was translated Finnish and back to English and approved by the AIDA research group.

Participants and procedure

The participants in this study were 402 high school students from four different schools in three different cities in Finland. During regular school hours, the teacher instructed students to complete all three measures anonymously using an internet-based form. Participation was voluntary, and all parents received an electronic information letter before their child participated. Students who did not want to participate were instructed to submit an empty form. The form also included a question of conscience at the end, where participants were asked if they answered the questions honestly. Before any statistical analyses, all participants who had sent an empty form (four) or who answered the question of conscience negatively (38) were removed from the data. The final sample included 360 students (14–21 years, $M = 16.39$). Of these participants, 228 (63.3%) were female and 132 were male (36.7%).

For CFA, a minimum requirement for sample size have been suggested to be 100-200, but adequate sample size is shown to be highly context-dependent (Wolf et al., 2013). We tried to maximize our limited resources for data collection and collect as much data as possible.

Statistical analyses

The Statistical Package for Social Sciences (SPSS 26 for Windows) was used for the analysis of descriptive statistics. We used independent sample t-tests to assess the differences between genders in the means of the different scales and subscales. The factor structure of the measures was assessed using CFA. In order to handle the variable non-normality, a robust maximum likelihood estimator was used. To assess missing data mechanism, we first applied Little's MCAR test. We further examined the missing data mechanism variable by variable by using t-tests to assess if respondents with missing data differ from respondents without missing data. The full information maximum likelihood method (FIML), without including auxiliary variables was used to handle the missing data, as suggested by Little

et al. (2014) when proportion of the missing data is not very high. All models were tested using the R environment for statistical computing (R Development Core Team, 2008) and the lavaan package (Rosseel, 2012).

For MZQ, we assessed two different measurement models. The authors of the measure (Hausberg et al., 2012) introduced a four-factor model but suggested using a one-factor MZQ before further validation. In line with their suggestion, we first used the one-factor model where all the items were set to load on a single factor. The second model included four factors. Because the two models were nested, we tested the chi-square difference of the models to compare their fitness to the data.

For DERS, we assessed four different measurement models. First, we considered the original full-length six-factor model suggested by Gratz and Roemer (2004). Second, we assessed the full-length five-factor DERS without the AWARENESS scale (DERS-R), a model suggested by Bardeen et al. (2012). Third, we looked at the shortened version of the original DERS (DERS-SF), which included six factors and was developed by Kaufman et al. (2016). Fourth, we evaluated a measurement model that included combined alterations to the original DERS; it was a shortened five-factor version of the DERS without the AWARENESS scale (DERS-R-SF).

For AIDA, we applied a higher order CFA using a measurement model consisting of two main scales (identity continuity vs. discontinuity and identity coherence vs. incoherence). Both higher order factors included three subscales. Continuity vs. discontinuity included the subscales *stability in attributes, stability in relations*, *and positive emotional self-reflection*. Coherence vs. incoherence included the subscales of *consistent self-image, autonomy*, and *positive cognitive self-reflection*. This model was the original AIDA model developed by Goth et al. (2012).

Because incremental fit indices compare the user measurement model against a supposedly poorly fitting baseline model, they might be misleading if the baseline model fits the data exceptionally well. It has been suggested that if the RMSEA of the baseline model is smaller than.158, the incremental goodness-of-fit indices may not be reliable (Kenny, 2015; Kenny et al., 2015). To address the issue with mixed results with goodness-of-fit indices of AIDA-models, we checked the root mean square error of approximation (RMSEA) of the baseline null model.

Four commonly recommended fit indices were used to evaluate the model fit: The Tucker-Lewis Index (TLI), the Comparative Fit Index (CFI), the RMSEA, and the Standardized Root Mean Square Residual (SRMR). The following guidelines were considered as an indication of a good fit: TLI close to .95, CFI close to .95, RMSEA close to .06 with 90% confidence interval with an upper limit $\leq$.08, and SRMR close to .08 (Hooper et al., 2008; Hu & Bentler, 1999).

Scale reliability was evaluated using the composite reliability coefficient omega (McDonald's ω), which is considered to provide a more accurate approximation of scale reliability than other measures (Revelle & Zinbarg, 2009). For an estimate of convergent validity, the value of the average variance extracted (AVE) was considered. To indicate an acceptable convergent validity, the AVE value should exceed .50. To assess the latent variable discriminant validity, we compared the AVE value to the shared variance of the latent variables. It is considered, that for any two constructs to indicate discriminant validity, the

AVE value of both constructs should be larger than the shared variance (squared correlation) of the two constructs (Fornell & Larcker, 1981).

Results

Descriptive statistics

The means and standard deviations, Cronbach's alphas, and ranges of corrected item-total correlations for MZQ, DERS, and DERS-SF total scales and subscales are summarized in Table 1. The mean scores for MZQ were lower for girls (indicating more reported mentalization difficulties) for all the scales except *refusing self-inspection*. For DERS, the mean scores were significantly higher for girls (indicating more reported ER difficulties) on all scales except for AWARENESS.

For AIDA, the proportion of missing data was totally 3.1%, ranging from 0.6 to 7.5% on individual variables. For DERS, range of missing data on single variable was 2.2-8.9%, and total proportion of missing data was 6.0%. For MZQ the data was missing in 8.4% of the datapoints, ranging from 5.3 to 10.3% on an individual variable. Little's MCAR test was insignificant for AIDA, giving us no reason to assume data from AIDA would not be missing completely at random (MCAR). For DERS and MZQ Little's MCAR test showed that missing data was not MCAR. However, the T-tests showed that for almost all variables for DERS and MZQ, respondents who had missing data did not differ from respondents without missing data. The only exceptions ($p < .05$) were items number 8 and 11 for MZQ, items 35 and 36 for DERS and item number 8 for AIDA. Therefore, concerning DERS and MZQ the data was assumed to be missing at random.

For DERS-SF, the AWARENESS and IMPULSE scores for boys and girls did not differ significantly; for all other scales, the girls had significantly higher scores than the boys. Correlations for DERS and the corresponding DERS-SF scales were all significant, with .84 for the total scale, .88 for AWARENESS, .91 for CLARITY, .96 for IMPULSE, .96 for GOALS, .97 for NONACCEPTANCE, and .93 for STRATEGIES.

The means and standard deviations, Cronbach's alphas, and ranges of corrected item total correlations for the AIDA main scales and the subscales are summarized in Table 2. For AIDA, the girls' mean scores were higher (indicating more identity problems) than the boys, but for the *stability in relations* subscale, the difference was not significant.

Mentalization

One-factor model

The one-factor measurement model for MZQ provided poor fit for the data: $\chi 2(90) = 317.69$, $p < .001$, CFI = .829, TLI = .801, RMSEA = .092 (90% confidence interval .082–.104), SRMR = .062. SRMS met the expected guideline, but other goodness-of-fit indices were quite far from expected guidelines. The composite reliability estimate for one factor was acceptable ($\omega = .87$), but the AVE value failed to show an indication of adequate convergent validity (AVE = .32).

Table 1. Descriptive statistics for MZQ, DERS and DERS-SF.

							Girls	Boys	
		N	M (SD)	α	Range of CITC	N	M (SD)	N	M (SD)
MZQ	Refusing	323	3.53 (0.85)	.55	.22-.44	202	3.54 (0.81)	121	3.52 (0.90)
	Awareness	313	3.49 (1.00)	.78	.47-.69	194	3.35 (1.01)	119	3.73 (0.96)*
	Equivalence	322	3.30 (0.96)	.68	.37-.55	199	3.13 (.092)	123	3.57 (0.98)*
	Regulation	326	3.73 (0.96)	.67	.32-.59	203	3.62 (0.97)	123	3.92 (0.90*
	Total	299	3.50 (0.80)	.88	.42-.69	184	3.40 (0.77)	115	3.67 (0.81)*
DERS	AWARENESS	323	2.54 (0.80)	.75	.36-.64	205	2.49 (0.77)	118	2.61 (0.83)
	CLARITY	339	2.08 (0.76)	.78	.45-.62	216	2.20 (0.78)	123	1.87 (0.66)*
	IMPULSE	316	1.98 (0.89)	.88	.33-.82	201	2.07 (0.93)	115	1.82 (0.79)*
	GOALS	324	2.83 (1.06)	.86	.41-.80	205	3.00 (1.06)	119	2.53 (0.99)*
	NONACCEPTANCE	322	2.03 (0.97)	.90	.60-.79	203	2.14 (0.98)	119	1.83 (0.95)*
	STRATEGIES	314	2.16 (0.91)	.90	.34-.76	199	2.30 (0.89)	115	1.92 (0.88)*
	Total	283	2.38 (0.72)	.94	-.10-.74	181	2.48 (0.72)	102	2.20 (0.68)*
DERS-	AWARENESS	339	2.16 (0.89)	.63	.36-.47	214	2.12 (0.85)	125	2.23 (0.95)
SF	CLARITY	342	1.93 (0.86)	.76	.53-.63	217	2.01 (0.85)	125	1.79 (0.87)*
	IMPULSE	324	1.82 (0.99)	.89	.76-.80	205	1.89 (1.02)	119	1.69 (0.93)
	GOALS	332	2.83 (1.23)	.89	.75-.82	211	3.00 (1.21)	121	2.53 (1.17)*
	NONACCEPTANCE	327	1.93 (1.02)	.85	.70-.74	207	2.02 (1.00)	120	1.77 (1.03)*
	STRATEGIES	328	1.96 (0.99)	.83	.68-.69	208	2.05 (1.00)	120	1.80 (0.94)*
	Total	309	2.09 (0.72)	.91	.13-.73	195	2.17 (0.71)	114	1.95 (0.72)*

MZQ, Mentalization Questionnaire; DERS, Difficulties in Emotion Regulation Scale; DERS-SF, Difficulties in Emotion Regulation Scale Short Form.
*Difference of means significant $<.05$.

Table 2. Descriptive statistics for AIDA.

						Girls		Boys
	N	M (SD)	α	Range of CITC	N	M (SD)	N	M (SD)
Discontinuity	308	1.24 (0.60)	.90	.27-.75	195	1.32 (0.60)	113	1.11 (0.59)*
Attributes	332	1.49 (0.65)	.74	.08-.57	214	1.55 (0.66)	118	1.39 (0.63)*
Relations	332	0.93 (0.61)	.79	.34-.61	212	0.96 (0.56)	120	1.86 (0.68)
Emotional	325	1.48 (0.91)	.83	.46-.65	206	1.56 (0.90)	121	1.15 (0.89)*
Incoherence	292	1.30 (0.74)	.95	.33-.74	184	1.44 (0.69)	108	1.05 (0.75)*
Consistent	319	1.39 (0.78)	.88	.25-.76	202	1.49 (0.81)	117	1.19 (0.84)*
Autonomy	321	1.33 (0.78)	.88	.34-.72	204	1.52 (0.72)	117	1.00 (0.78)*
Cognitive	332	1.18 (0.78)	.83	.45-.64	209	1.28 (0.73)	123	1.02 (0.83)*
Total	267	1.28 (0.65)	.96	.26-.76	169	1.40 (0.63)	98	1.09 (0.64)*

AIDA, Assessment of Identity Development in Adolescence.
*Difference of means significant $<.05$.

Table 3. MZQ one factor model and 4 factor model composite reliability and average variance extracted.

Factor	ω	AVE
One factor model	.86	.32
1. Refusing	.53	.23
2. Awareness	.79	.48
3. Equivalence	.71	.38
4. Regulation	.73	.47
Total	.86	.38

MZQ, Mentalization Questionnaire.

Four-factor model

The original four-factor measurement model suggested by Hausberg et al. (2012) provided a significantly better fit for the data than the one-factor model: $\chi 2$ (6) = 35.96, $p < .001$. However, even though the four-factor model fit the data better than the one-factor model, the overall fit of the four-factor model was not adequate: $\chi 2(84) = 271.85$, $p < .001$, CFI = .865, TLI = .831, RMSEA = .085 (90% confidence interval .074–.096), SRMR = .063. SRMS met the expected guideline, and RMSEA was close to the guideline, but CFI and TLI did not indicate a good fit.

Composite reliability was acceptable for all factors except *refusing self-inspection*. AVE values, however, did not indicate adequate convergent validity for any factor or for the overall scale (see Table 3). Comparisons between squared correlations and AVE values failed to indicate discriminant validity between factors, as all the AVE values were lower than the lowest shared variance (squared correlation) between the constructs.

Exploratory model

As neither model was adequately confirmed, we continued using CFA with modification indices and by allowing cross loadings to explore an alternative measurement model that would fit the data better. We examined from modification indices which parameters seemed to cause largest model misfit and allowed one item at time to load on two factors, the factor suggested by original model and the factor suggested by the modification indices. When the item seemed to load clearly better on the alternative factor, we re-specified the model accordingly and checked the model fit and the modification indices after the modification and repeated these steps, until the model fit was acceptable. This step-by-step approach led us to move item 9 from *refusing self-reflection* to *emotional awareness* -factor, item 15 from *emotional awareness* to *regulation of affect* -factor, and item 2 from *regulation of affect*, items 4 and 7 from *psychic equivalence mode* and item 8 from *emotional awareness* to *refusing self-reflection* -factor. We renamed *emotional awareness* as F1, *psychic equivalence mode* as F2, *regulation of affect* as F3 and *refusing self-reflection* as F4 (see Table 5). The alternative four-factor exploratory model fit the data well, with $\chi 2(84) = 171.22$, $p < .001$, CFI = .950, TLI = .930, RMSEA = .057 (90% confidence interval .045–.070), and SRSM = .043. All goodness-of-fit indices were fulfilled or were close to expected guidelines.

All factors showed acceptable reliability. The values for AVE indicated a convergent validity for factors F1, F2, and F3, but was lower than expected for F4. Total AVE was slightly

Table 4. MZQ exploratory model composite reliability, average variance extracted and range of corrected item-total correlations.

Factor	ω	AVE	range of CITC
F1	.76	.53	.49-.64
F2	.77	.63	.61
F3	.81	.59	.59-.68
F4	.76	.32	.38-.54
Total	.90	.46	

MZQ, Mentalization Questionnaire.

Table 5. Standardized factor loadings for MZQ exploratory model.

Item #	Factors			
	F1	F2	F3	F4
9	.55			
10	.75			
11	.86			
1		.74		
12		.84		
3			.78	
6			.75	
15			.77	
2				.52
4				.50
5				.57
7				.65
8				.66
13				.52
14				.55

MZQ, Mentalization Questionnaire.

lower than expected (see Table 4). Factor correlations were all significant and ranged from 0.56 to 0.82. Comparisons between AVE and squared correlations indicated discriminant validity for F1 from F2, and F2 from F3. Single item factor loadings ranged from .55 to .86 for F1, from .74 to .84 for F2, from .75 to .78 for F3, and from .50 to .66 for F4 (see Table 5).

Emotion regulation

Full-length DERS

The full-length DERS model with six factors did provide a somewhat acceptable, but not good fit for the data: $\chi 2(579) = 1578.62$, $p < .001$, CFI = .844, TLI = .830, RMSEA = .076 (90% confidence interval .071–.080), SRSM = .060. SRSM met the guideline for model fit, and

Table 6. Composite reliability and average variance extracted DERS/DERS-R/DERS-SF/DERS-R-SF.

Factor	ω	AVE
1. AWARENESS	.76/ - /.64/ -	.35/ - /.37/ -
2. CLARITY	.78/.77/.76/.76	.42/.41/.51/.51
3. IMPULSE	.88/.88/.89/.89	.58/.58/.72/.72
4. GOALS	.87/.88/.89/.89	.60/.60/.73/.73
5. NONACCEPTANCE	.91/.91/.84/.84	.61/.61/.64/.64
6. STRATEGIES	.89/.89/.83/.83	.54/54/.62/.62
Total	.93/.95/.94/.95	.52/.56/.61/.65

DERS, Difficulties in Emotion Regulation Scale; DERS-R, Difficulties in Emotion Regulation Scale Revised; DERS-SF, Difficulties in Emotion Regulation Scale Short Form.

RMSEA was close to the guideline, but other goodness-of-fit indices failed to show adequate fit to the data.

All factors showed acceptable reliability. The values for AVE failed to show an indication of convergent validity for AWARENESS and CLARITY, but the values for IMPULSE, GOALS, NONACCEPTANCE, STRATEGIES, and total AVE were acceptable (see Table 6). Factor correlations ranged from −.05 to .82. AWARENESS correlated significantly only with the CLARITY factor. For the other factors, intercorrelations were high. Comparisons between AVE values and squared correlations between factors indicated discriminant validity for AWARENESS from all factors, and for CLARITY from GOALS, IMPULSE from GOALS and NONACCEPTANCE, and for GOALS from NONACCEPTANCE. Single item factor loadings ranged from .44 to .76 for AWARENESS, from .53 to .70 for CLARITY, from .33 to .86 for IMPULSE, from .42 to .87 for GOALS, from .66 to .84 for NONACCEPTANCE, and from .36 to .80 for STRATEGIES (see Table 7).

DERS without AWARENESS-factor (DERS-R)

The measurement model of the full-length DERS with five factors did provide a somewhat acceptable, but not good, fit for the data: $\chi 2(395) = 1116.25$, $p < .001$, CFI = .869, TLI = .855, RMSEA = .080 (90% confidence interval .075–.086), SRSM = .070. SRSM met the expected guideline, but other goodness-of-fit indices did not.

Indication for overall reliability was good, with $\omega = .96$. Indication of convergent validity was shown for the overall scale and for all other factors except CLARITY (see Table 6). Correlations between factors ranged from .57 to .82, and comparisons of AVE values and squared correlations between factors indicated discriminant validity for GOALS from CLARITY and for IMPULSE from NONACCEPTANCE. Factor loadings for single items ranged from .48 to .73 for CLARITY, from .33 to .86 for IMPULSE, from .43 to .86 for GOALS, from .65 to .84 for NONACCEPTANCE, and from .36 to .80 for STRATEGIES (see Table 7).

Short version of DERS (DERS-SF)

The measurement model for the short version of DERS provided a good fit for the data: $\chi 2(120) = 186.62$, $p < .001$, CFI = .974, TLI = .966, RMSEA = .044 (90% confidence interval .031–.056), SRSM = .036. All goodness-of-fit indices met the expected guidelines.

Table 7. Standardized factor loadings from confirmatory factor analyses of different DERS versions.

Item#	Factors: DERS / DERS-R / DERS-SF / DERS-R-SF					
	1	2	3	4	5	6
2R	.65/ - /.68/ -					
6R	.76/ - / - / -					
8R	.71/ - /.66/ -					
10R	.45/ - /.48/ -					
17R	.50/ - / - / -					
34R	.44/ - / - / -					
1R		.64/.61/ - / -				
4		.70/.72/.71/.71				
5		.70/.73/.73/.73				
7		.53/.48/- / -				
9		.65/.67/.71/.71				
3			.78/.78/ - / -			
14			.85/.85/.83/.83			
19			.85/.85/ - / -			
24R			.33/.33/ - / -			
27			.86/.86/.87/.87			
32			.84/.84/.86/.86			
13				.82/.82/.81/.81		
18				.87/.86/.89/.89		
20R				.42/.43/- / -		
26				.86/.86/.86/.86		
33				.75/.75/ - / -		
11					.77/.77/ - / -	
12					.82/.82/.78/.77	
21					.84/.84/ - / -	
23					.66/.65/ - / -	
25					.81/.81/.83/.84	
29					.80/.80/.79/.79	
15						.80/.80/ - / -
16						.75/.75/.78/.78
22R						.36/.36/ - / -
28						.76/.76/.79/.79
30						.77/.77/ - / -
31						.76/.76/ - / -
35						.80/.80/.79/.79
36						.80/.80/ - / -

Factors: 1 = AWARENESS, 2 = CLARITY, 3 = IMPULSE, 4 = GOALS, 5 = NONACCEPTANCE, 6 = STRATEGIES; R = reversed item; DERS, Difficulties in Emotion Regulation Scale; DERS-R, Difficulties in Emotion Regulation Scale Revised; DERS-SF, Difficulties in Emotion Regulation Scale Short Form.

The DERS-SF AWARENESS factor failed to show adequate reliability, but for all other factors, composite reliability was acceptable (see Table 6). AVE values showed an indication of convergent validity for CLARITY, IMPULSE, GOALS, NONACCEPTANCE, and STRATEGIES

Table 8. Composite reliability and average variance extracted for AIDA subscales.

Factor	ω	AVE
1.1 Attributes	.76	.27
1.2 Relations	.73	.27
1.3 Emotional	.86	.47
2.1 Consistant	.90	.44
2.2 Autonomy	.89	.40
2.3 Cognitive	.83	.39
Total	.95	.38

AIDA, Assessment of Identity Development in Adolescence.

factors and for the total scale, but for the AWARENESS factor, the AVE value was lower than expected. Correlations between factors ranged from .03 to .79. As in the full-length version, correlations between AWARENESS and other factors were substantially lower than the intercorrelations between other factors. AWARENESS and GOALS factors did not correlate with each other. Comparisons between AVE values and squared correlations indicated discriminant validity for AWARENESS from all other factors, for GOALS from CLARITY, IMPULSE, NONACCEPTANCE, and STRATEGIES, and NONACCEPTANCE from CLARITY and IMPULSE, and IMPULSE from STRATEGIES. Factor loadings ranged from .48 to .68 for AWARENESS, from .70 to .73 for CLARITY, from .83 to .87 for IMPULSE, from .81 to .89 for GOALS, from .77 to .83 for NONACCEPTANCE, and from .78 to .79 for STRATEGIES (see Table 7).

Short version of DERS without AWARENESS-factor (DERS-R-SF)

A measurement model for DERS-R-SF fit the data well: $\chi 2(80) = 144.40$, $p < .001$, CFI = .971, TLI = .962, RMSEA = .055 (90% confidence interval.041–.070), SRMR = .034. All goodness-of-fit indices met the expected guidelines. All factors indicated acceptable reliability, and AVE values showed an indication for convergent validity for all factors (see Table 6). Correlations between factors ranged from .52 to .79, and comparisons between AVE values and squared correlations showed indication for discriminant validity for CLARITY from GOALS and NONACCEPTANCE, for IMPULSE from GOALS, NONACCEPTANCE and STRATEGIES, and for GOALS from NONACCEPTANCE and STRATEGIES. Factor loadings for all items were above .71 (see Table 7).

Identity integration

Original AIDA-model

The fit indices for higher order CFA measurement model for AIDA showed mixed results: $\chi 2(1588) = 3722.45$, $p < .001$, CFI = .749, TLI = .739, RMSEA = .065 (90% confidence interval .062–.068), SRMR = .081. RMSEA and SRMR met the expected guidelines, but incremental goodness-of-fit indices (CFI and TLI) indicated a poor fit for the model. The RMSEA of the baseline null model was .129. This suggests that the incremental goodness-of-fit indices

Table 9.. AIDA explorative model.

Scale	Items
Discontinuity	
Relations	1R, 2R, 5R, 8, 9, 10, 18, 23R, 28, 39R, 40, 41R, 43R, 54, 55, 58R
Emotional	3, 11, 17R, 19, 24, 26R, 27, 29, 30, 33R, 44
Incoherence	
Consistent	4, 12, 13, 15, 25, 31, 32, 45, 47, 56R, 57
Autonomy	142,021,223,436,384,246,485,053
Cognitive	6,71,63,53,74,95,152
Method factor	1R, 2R, 5R, 17R, 23R, 26R, 33R, 39R, 41R, 43R, 56R, 58R

AIDA, Assessment of Identity Development in Adolescence; R = reverse coding item.

would not be useful; absolute indices would be more credible, and the model would fit the data.

All factors showed an indication for acceptable reliability, but none of the AVE values showed an indication for convergent validity for the subscales (see Table 8). The correlation between the two higher order scales was significant (.96). The correlations between the lower order subfactors were all significant ranging from .56 to .82. Comparisons between AVE values and squared correlations failed to indicate discriminant validity between the subscales. The standardized factor loadings for the lower-order subscales ranged from .10 to .67 for *stability in attributes*, from .34 to .80 for *stability in relations*, from .51 to .77 for *positive emotional self-reflection*, from .34 to .79 for *consistent self-image*, from .37 to .75 for *autonomy*, and from .50 to .70 for *positive self-reflection*. For higher-order factors, the standardized loadings of second-order factors were .62 to .97 for discontinuity and from .93 to .98 for incoherence.

Explorative model with a method factor

Even if the original model fit the data, the fact that the first subscale (*stability in attributes*) consisted of many of the reverse-coded items raised the possibility that the subscale would have features of a method factor. For evaluating possible method factor effects, we constructed an explorative alternative model for testing. The explorative model included two main scales like the original model, but the first subscale (*stability in attributes*) was removed, and the corresponding items were divided into two subscales on the continuity main scale. This was done by allowing individual items to cross load on both subscales and reviewing from factor loadings and modification indices on which of the subscales seemed more appropriate for individual item. Since the AIDA model is a higher order model, where all the items on *stability of attributes* -subscale also belong on the higher order main scale (continuity), we found it also theoretically justifiable to assume that model could be specified with these items on different subscales on the continuity main scale. The fact that subscales correlate highly with each other also affirms this. The explorative model also incorporated a method factor that included all reverse-coded items. For the method factor,

correlation with other factors was set to zero. Details of the individual items on different scales and subscales of the explorative model are shown on Table 9.

We tested the chi-square differences between the two models to compare their fit to the data. Goodness-of fit indices showed mixed results for the explorative model: $\chi2(1577) = 3375.38$, $p < .001$, CFI = .789, TLI = .778, RMSEA = .060 (90% confidence interval .057–.063), SRMR = .060. The RMSEA of the baseline model (.129) favors the use of absolute fit indices and accepting the model. The explorative model with the method factor fit the data significantly better than the original model: $\chi2\ (11) = 359.71$, $p < .001$.

Discussion

This study aimed to evaluate the construct validity and reliability of measures for mentalization, ER difficulties, and identity diffusion when used in a normative sample of Finnish adolescents. Finding adequate tools to measure these transdiagnostic constructs will have a great clinical importance. Paradigm for assessing and diagnosing personality disorders has lately been shifting from categorical to dimensional approach (Mulder & Tyrer, 2019). BPD has been suggested to capture impairment in personality functioning in more general (Goth et al., 2012; Mulder & Tyrer, 2019), and reliable and valid measures of concepts central in different treatments for BPD have potential for use in clinical practice for assessing features of personality functioning beyond symptoms and categorical diagnoses for improving functionality of assessment in planning and monitoring treatment, as well as bringing scientific and clinical, and psychological and psychiatric points of view more together with each other.

Mentalization

The original validation for MZQ was conducted in an adult clinical sample, but our results were similar to a nonclinical sample of Italian adolescents (Belvederi Murri et al., 2017); the data mean scores and standard deviations in the two studies were similar. In our sample, the female participants reported more difficulty with mentalization overall and more difficulty on the subscales than the male participants. Previous studies have not reported data separately for males and females, so we cannot compare our data. Differences between genders might reflect that MZQ captures difficulties in mentalization more typical for females than for males, for example momentary failures resulting from intensive emotions, and that difficulties more typical for males would not be so well recognized with MZQ. Since the females in our sample, as in previous studies, reported more difficulty in ER and identity, it seems plausible that MZQ grasps in particular such difficulties in Mentalization which correlate with problems in ER and identity. Previous studies in general population using observer-related measures such as RF scale, have reported that adult women seem to have higher levels of mentalization compared on men (Jessee et al., 2016). Differences between genders among adolescents have not been found (Chow et al., 2017). These findings would suggest that MZQ indeed captures different aspects of mentalization than RF scale, or that girls would be more aware of their problems with mentalization than boys in our sample. This highlights the need for MZQ validation studies against observer rated

measures, as suggested by scale authors in their original validation study (Hausberg et al., 2012).

Hausberg et al. (2012) suggested using the whole MZQ scale before further validation of the measure. In our results, however, the one-factor model had a poor fit, and the AVE value of the one-factor model indicated that most of the variance explained by the model was due to measurement error. These results contrast with the use of the total score of the measure as a valid indication of mentalization ability.

The four-factor measurement model of MZQ did not provide a very good fit with the data. The reliability of the overall scale, as well as subscales other than *refusing self-inspection*, was acceptable. Our findings regarding subscale reliability are in line with Hausberg et al. (2012). In our findings on the factor structure of the measure, SRMS showed similar results, but the RMSEA indicated a poorer fit on our CFA. In contrast to Hausberg et al. (2012), we used incremental fit indices in addition to absolute fit indices to assess the model fit, thus obtaining a more multifaceted view of the fit.

In the original validation study, Hausberg et al. (2012) described the indication of convergent and discriminant validity for the MZQ total score assessed with correlations with measures of symptom severity. We used AVE values to assess the convergent validity of the subscales, and we also assessed the discriminant validity of the subscales by comparing AVE values and factor correlations. In our findings, lower than expected AVE values suggested problems with the convergent validity of the subscales. AVE represents the average amount of variance that the construct explains in its indicator variables relative to the overall variance of its indicators. None of the AVE values in our MZQ measurement models reached a satisfactory level, implying that most of the variance explained by the latent variables of the measurement models reflect measurement error. Also, due to low AVE values and high correlations between factors, our data suggest that the discriminant validity of factors proposed by Hausberg et al. (2012) is low. According to the indications of unsatisfactory convergent and discriminant validity, there are many items with quite low ($<.50$) factor loadings, implying item cross loadings.

It is possible that the factor structure and validity of the measure was somewhat different in our data than in the original validation sample because we used a population sample instead of a clinical sample. It is also possible that our results reflect an unsuccessful Finnish translation of the MZQ and that the items in the English-language version capture the true mentalization ability with less measurement error. However, there is a need for further studies regarding MZQ before recommending it as a valid measure of adolescent mentalization ability among Finnish adolescents.

Finally, we used modification indices to find an alternative exploratory model for MZQ after finding the initial model did not fit the data well. In our model, F1 can be considered for factor measuring ER and the understanding of one's own emotions. F2 can be considered as a factor that relates to perceiving relationships and being criticized as a major threat. F3 seems to be related to perceiving emotions as a threat and an uncontrollable force, and this feature could be related to the risk for mentalization failures caused by intense emotions. In our model, F4 consists of items reflecting rigid and concrete thinking and a nonmentalizing attitude toward one's mind in general. However, because modification indices can lead to model overfitting, it is important to note that this explorative model

may fit only this data. Using these results in a more general way requires replicating it in other data.

Emotion regulation

We assessed four different measurement models for the versions of DERS. Our findings are in line with Kaufman et al. (2016), indicating that the short version of DERS shows similar, and partly stronger, psychometric properties than the full version, at least among adolescents, and our data gives more weight to recommendations suggesting the use of the shorter version of the measure. Reliability seemed to be very similar between the short version (without items with weak factor loadings) and the full version, but the AVE values (an indication of convergent validity) were generally stronger in the short version.

Bardeen et al. (2012) pointed out that the AWARENESS subscale does not correlate well with other subscales and does not contribute much to the general DERS factor. They have suggested that the AWARENESS subscale does not belong to the same higher-order construction as other subscales and recommended using DERS without the AWARENESS subscale as a more contiguous, unified measure. Our data is in line with their findings. AWARENESS is also the only DERS subscale using mostly reversed items, and these items may create a method factor that reflects more effects from the measure than the effects from the measured psychological construct (DiStefano & Motl, 2006).

There have been other suggestions about handling the problems of the AWARENESS subscale. Cho and Hong (2013) have suggested that the AWARENESS and CLARITY subscales could be combined into one, as an *understanding emotions* construct with a controlled method factor for reverse-scored items. However, in our data, the AWARENESS subscale was the psychometrically weakest, with the lowest reliability, convergent validity (AVE), and item factor loadings in the short and full versions of the measure. This suggests that removing the factor from the measure would be the simplest and, perhaps, most recommendable way of using the DERS.

In our sample, females reported more overall ER difficulties than males. On the AWARENESS subscale, however, the mean score for boys was higher than for girls, but the difference was not statistically significant. In the short version of the scale, the results were otherwise similar, but in the IMPULSE subscale, the difference between the genders was not statistically significant. In a previous general population validation study with Dutch adolescents (Neumann et al., 2010), girls scored significantly higher than boys on total scores and on most subscales. In the results for AWARENESS, boys had higher scores than girls; on the IMPULSE subscale, there were no gender differences in their sample. Overall, the gender differences in our sample were very similar to a study by Neumann et al. (2010), even though there were slight differences between the samples.

Identity integration

We received mixed CFA results concerning the fit of the data using the AIDA model. The absolute goodness-of fit indices (RMSEA, SRMR, and $\chi 2/df$) indicated acceptable fit for the model, but the incremental goodness-of-fit indices (CFI and TLI) were quite far from the expected guidelines, suggesting a poor fit. However, as incremental fit indices compare the user measurement model against a supposedly poorly fitting baseline model, they might be

misleading if the baseline model fits the data exceptionally well (Kenny, 2015; Kenny et al., 2015). In our study, this was the case. The RMSEA of the baseline model suggested that the absolute goodness-of-fit indices would be more credible, and the construct of the AIDA measurement model developed by Goth et al. (2012) would be convergent with our data.

The composite reliability of the AIDA subscales was acceptable, and the estimates of the reliability were very similar to Goth et al. (2012) original validation study. The AVE values of the subscales, however, were lower than expected, implying that AIDA produces a significant amount of measurement error and that the items inside the subscales do not necessarily measure the same things. Low AVE values also resulted in a lack of discriminant validity between the subscales. One possible way to improve the measure could be the removal of some items with low factor loadings to examine whether the coherence of the subscales would improve without jeopardizing reliability.

Another possible problem of the measure could be that many reverse coding items seem to cluster on the first factor (*stability in attributes*). This suggests that the stability in attributes subscale has features of a method factor, reflecting the properties of the measure and the differences or biases in the answering style of the respondents rather than the differences in the underlying psychological construct. To evaluate this issue, we built a model where all reversed items were controlled as a method factor and all items on the first factor were collapsed into other subfactors with the same main factor. It fitted the data better than the original model, suggesting that the reverse-coded items affected the factor structure of the measure. However, more confirmative studies with different datasets are needed.

The bias with reverse coding items should not have a great impact on the clinical usefulness of the measure because the higher factor structure of the measure seems to be confirmed and all the factors correlate with each other quite strongly. However, with scientific use, it is important to be aware of this issue because the method factor might affect, for example, the fit of some more complex structural equation models. The debate regarding the pros and cons of reverse coding items is ongoing. Some researchers suggest that reverse coding items should not be used at all, while others see them as useful despite possible bias with the factor structures. Weijters et al. (2013) have stated that the same biases affecting reverse coding items also partly affect direct items, and using reverse items at least brings these biases out for us to acknowledge.

In our sample, the girls reported more problems with identity than the boys overall. The differences between the genders were also significant on the subscales, except for *stability in relations*. Our findings regarding gender are quite similar to the original validation study of the measure (Goth et al., 2012), where girls had slightly higher mean scores than boys on all scales, but on the *stability in the attributes* subscale the gender difference was not significant.

Limitations

This study has some limitations. Our data does not include any criterion variables, such as alternative measures of mentalization, ER, identity, or measures of the mental health variables. This limitation prevents us from addressing the criterion validity of the measures or the convergent validity of the full measures. Our scope, therefore, is limited to assessing

the factorial structure of the measures, the convergent and discriminative validity, and the coherence of the different subscales. Our study used a general adolescent population data, and that limits generalizing our findings directly to specific patient populations. Data was collected from four schools in three different cities. Schools were selected by contacting larger number of schools and including them who were willing to participate. This type of selection has potential for biasing results, but for the schools included and schools in Finland in general are quite homogenous, bias in unlikely.

Adequate sample size in CFA is highly contextual (Wolf et al., 2013), and the best way to determine adequate sample size would be by conducting simulations prior data collection. We were unable to do this, and our sample size is based on more uncertain rules of thumb. Therefore, our sample size might be problematic, mainly for AIDA models with relatively high number of items, and parameters to be estimated. However, in our results regarding AIDA models, there were no straightforward indications of limited statistical power, e.g. insignificant parameters.

Conclusions

The results of the present study indicate that there is still some work to do in order to gain the valid measurement of three core transdiagnostic concepts, mentalization, ER and identity diffusion. First, the factor structure of the Finnish translation of MZQ was not confirmed for use among adolescents, and the measure seemed to produce quite a significant amount of measurement error. There is a need for further studies concerning the structure and construct validity of the MZQ. Second, the results from this study imply that some modification may be necessary in order to obtain the most reliable results from DERS and AIDA. Especially, we recommend using the short version of DERS without the *lack of emotional awareness* subscale. Concerning AIDA, our results confirmed the overall factor structure, but the *stability in the attributes* subscale might have some features of a method factor, and this should be taken into account when using AIDA for research. Also, the subscales show a somewhat problematic amount of measurement error affecting the convergent and discriminative validity of the subscales.

All procedures performed in studies involving human participants were in accordance with the ethical standards of the Ethics Committee of the Tampere region (31/2017), and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Informed consent was obtained from all individual participants included in the study. The authors declare that they have no conflict of interest.

Confict of interest

The authors declare that they have no conflict of interest.

Funding

This study has not received any external funding.

Author note

Parts of data analyses considering DERS and MZQ have been presented at Society for Psychotherapy Research 50th International Annual Meeting in Buenos Aires, Argentina (2019). All correspondence conserning this article should be addressed to Sami J. Eloranta, Nuorisopsykiatrian poliklinikka, Tampere University Hospital, PL 20,00,33,521 Tampere, Finland. Email: sami.eloranta@tuni.fi https://orcid.org/0000-0001-8691-6957

ORCID

Sami J. Eloranta http://orcid.org/0000-0001-8691-6957

REFERENCES

Atkinson, T. M., Rosenfeld, B. D., Sit, L., Mendoza, T. R., Fruscione, M., Lavene, D., Shaw, M., Li, Y., Hay, J., Cleeland, C. S., Scher, H. I., Breitbart, W. S., & Basch, E. (2011). Using confirmatory factor analysis to evaluate construct validity of the Brief Pain Inventory (BPI). *Journal of Pain and Symptom Management*, *41*(3), 558–565. https://doi.org/10.1016/j.jpainsymman.2010.05.008

Bardeen, J. R., Fergus, T. A., & Orcutt, H. K. (2012). An examination of the latent structure of the Difficulties in Emotion Regulation Scale. *Journal of Psychopathology and Behavioral Assessment*, *34*(3), 382–392. https://doi.org/10.1007/s10862-012-9280-y

Belvederi Murri, M., Ferrigno, G., Penati, S., Muzio, C., Piccinini, G., Innamorati, M., Ricci, F., Pompili, M., & Amore, M. (2017). Mentalization and depressive symptoms in a clinical sample of adolescents and young adults. *Child and Adolescent Mental Health*, *22*(2), 69–76. https://doi.org/10.1111/camh.12195

Bowlby, J. (1988). *A secure base*. Routledge.

Calkins, S. D., & Hill, A. (2007). Caregiver influences on emerging emotion regulation: Biological and environmental transactions in early development. In J. J. Gross (Ed.) *Handbook of emotion regulation*. Guilford Press.

Camoirano, A. (2017). Mentalizing makes parenting work: A review about parental reflective functioning and clinical interventions to improve it. *Frontiers in Psychology*, *8*, 14. https://doi.org/10.3389/fpsyg.2017.00014

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105. https://doi.org/10.1037/h0046016

Cho, Y., & Hong, S. (2013). The new factor structure of the Korean version of the difficulties in emotion regulation scale (K-DERS) incorporating method factor. *Measurement and Evaluation in Counseling and Development*, *46*(3), 192–201. https://doi.org/10.1177/0748175613484033

Chow, C. C., Nolte, T., Cohen, D., Fearon, R. M. P., & Shmueli-Goetz, Y. (2017). Reflective functioning and adolescent psychological adaptation: The validity of the reflective functioning scale-adolescent version. *Psychoanalytic Psychology*, *34*(4), 404–413. https://doi.org/10.1037/pap0000148

DiStefano, C., & Motl, R. W. (2006). Further investigating method effects associated with negatively worded items on self-report surveys. *Structural Equation Modeling: A Multidisciplinary Journal*, *13*(3), 440–464. https://doi.org/10.1207/s15328007sem1303_6

Erikson, E. H. (1968). *Identity: Youth and crisis*. W. W. Norton.

Fonagy, P. (2000). Attachment and borderline personality disorder. *Journal of the American Psychoanalytic Association*, *48*(4), 1129–1146. https://doi.org/10.1177/00030651000480040701

Fonagy, P., & Bateman, A. (2008). The development of borderline personality disorder-a mentalizing model. *Journal of Personality Disorders*, *22*(1), 4–21. https://doi.org/10.1521/pedi.2008.22.1.4

Fonagy, P., Bateman, A., & Bateman, A. (2011). The widening scope of mentalizing: A discussion. *Psychology and Psychotherapy: Theory, Research and Practice*, *84*(1), 98–110. https://doi.org/10.1111/j.2044-8341.2010.02005.x

Fonagy, P., Gergely, G., Jurist, E. L., & Target, M. (2002). *Affect regulation, mentalization and the development of the self*. Other Press.

Fonagy, P., Target, M., Steele, M., & Steele, H. (1998). *Reflective-Functioning Manual version 5.0, for application to adult attachment interviews*. University College.

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, *18*(1), 39–50. https://doi.org/10.2307/3151312

Freeman, C. (2016). What is mentalizing? An overview. *British Journal of Psychotherapy*, *32*(2), 189–201. https://doi.org/10.1111/bjp.12220

Glenn, C. R., & Klonsky, E. D. (2009). Emotion dysregulation as a core feature of borderline personality disorder. *Journal of Personality Disorders*, *23*(1), 20–28. https://doi.org/10.1521/pedi.2009.23.1.20

Goth, K., Foelsch, P., Schlüter-Müller, S., Birkhölzer, M., Jung, E., Pick, O., & Schmeck, K. (2012). Assessment of identity development and identity diffusion in adolescence – Theoretical basis and psychometric properties of the self-report questionnaire AIDA. *Child and Adolescent Psychiatry and Mental Health*, *6*(1), 27. https://doi.org/10.1186/1753-2000-6-27

Gratz, K. L., & Roemer, L. (2004). Multidimensional assessment of emotion regulation and dysregulation: Development, factor structure, and initial validation of the difficulties in emotion regulation scale. *Journal of Psychopathology and Behavioral Assessment*, *26*(1), 41–54. https://doi.org/10.1023/B:JOBA.0000007455.08539.94

Gross, J. J., & Thompson, R. A. (2007). Emotion regulation: Conceptual foundations. In J. J. Gross (Ed.) *Handbook of emotion regulation* (pp. 3–24). Guilford Press.

Hausberg, M. C., Schulz, H., Piegler, T., Happach, C. G., Klöpper, M., Brütt, A. L., Sammet, I., & Andreas, S. (2012). Is a self-rated instrument appropriate to assess mentalization in patients with mental disorders? Development and first validation of the Mentalization Questionnaire (MZQ). *Psychotherapy Research*, *22*(6), 699–709. https://doi.org/10.1080/10503307.2012.709325

Hooper, D., Coughlan, J., & Mullen, M. (2008). Structural equation modelling: Guidelines for determining model fit. *Electronic Journal of Business REsearch Methods*, *6*(1), 53–59. https://doi.org/10.1037/1082-989X.12.1.58

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

Jessee, A., Mangelsdorf, S. C., Wong, M. S., Schoppe-Sullivan, S. J., & Brown, G. L. (2016). Structure of reflective functioning and adult attachment scales: Overlap and distinctions. *Attachment & Human Development*, *18*(2), 176–187. https://doi.org/10.1080/14616734.2015.1132240

John, O., & Benet-Martínez, V. (2000). Measurement: Reliability, construct validation, and scale construction. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 339–369). Cambridge University Press. https://psycnet.apa.org/record/2014-12227-018

Jung, E., Pick, O., Schlüter-Müller, S., Schmeck, K., & Goth, K. (2013). Identity development in adolescents with mental problems. *Child and Adolescent Psychiatry and Mental Health*, *7*(1), 26. https://doi.org/10.1186/1753-2000-7-26

Katznelson, H. (2014). Reflective functioning: A review. *Clinical Psychology Review*, *34*(2), 107–117. https://doi.org/10.1016/j.cpr.2013.12.003

Kaufman, E. A., Xia, M., Fosco, G., Yaptangco, M., Skidmore, C. R., & Crowell, S. E. (2016). The difficulties in emotion regulation scale short form (DERS-SF): Validation and replication in adolescent and adult samples. *Journal of Psychopathology and Behavioral Assessment*, *38*(3), 443–455. https://doi.org/10.1007/s10862-015-9529-3

Keinänen, M., Martin, M., & Lindfors, O. (2017). Ryhmäterapia ja mentalisaatio. In I. Laitinen & S. Ollikainen (Eds.) *Mentalisaatio: Teoriasta Käytäntöön* (pp. 181–202). Therapeia-säätiö.

Keinänen, M., Martin, M., & Lindfors, O. (2019, September 30). MZQ, Mentalization Questionnaire (Trans). https://ammattilaiset.mielenterveystalo.fi/tyokalut/mittaripankki

Kenny, D. A. (2015). Measuring model fit. Retrieved August 2, 2019, from http://davidakenny.net/cm/fit.htm

Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2015). The performance of RMSEA in models with small degrees of freedom. *Sociological Methods & Research*, *44*(3), 486–507. https://doi.org/10.1177/0049124114543236

Kernberg, O. F. (2006). Identity: Recent findings and clinical implications. *The Psychoanalytic Quarterly*, *75*(4), 969–1004. https://doi.org/10.1002/j.2167-4086.2006.tb00065.x

Little, T. D., Jorgensen, T. D., Lang, K. M., & Moore, E. W. G. (2014). On the joys of missing data. *Journal of Pediatric Psychology*, *39*(2), 151–162. https://doi.org/10.1093/jpepsy/jst048

McArdle, J. J. (1996). Current directions in structural factor analysis. *Current Directions in Psychological Science*, *5*(1), 11–18. https://doi.org/10.1111/1467-8721.ep10772681

McLaughlin, K. A., Hatzenbuehler, M. L., Mennin, D. S., & Nolen-Hoeksema, S. (2011). Emotion dysregulation and adolescent psychopathology: A prospective study. *Behaviour Research and Therapy*, *49*(9), 544–554. https://doi.org/10.1016/j.brat.2011.06.003

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

Neumann, A., van Lier, P. A. C., Gratz, K. L., & Koot, H. M. (2010). Multidimensional assessment of emotion regulation difficulties in adolescents using the difficulties in emotion regulation scale. *Assessment*, *17*(1), 138–149. https://doi.org/10.1177/1073191109349579

R Development Core Team. (2008). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.r-project.org

Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika*, *74*(1), 145–154. https://doi.org/10.1007/s11336-008-9102-z

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. R package version 0.5-15 http://lavaan.org. *Journal of Statistical Software*, *48*(2), 1–36. https://users.ugent.be/~yrosseel/lavaan/lavaanIntroduction.pdf https://doi.org/10.18637/jss.v048.i02

Schwartz, S. J., Hardy, S. A., Zamboanga, B. L., Meca, A., Waterman, A. S., Picariello, S., Luyckx, K., Crocetti, E., Kim, S. Y., Brittian, A. S., Roberts, S. E., Whitbourne, S. K., Ritchie, R. A., Brown, E. J., & Forthun, L. F. (2015). Identity in young adulthood: Links with mental health and risky behavior. *Journal of Applied Developmental Psychology*, *36*, 39–52. https://doi.org/10.1016/j.appdev.2014.10.001

Sollberger, D. (2013). On identity: From a philosophical point of view. *Child and Adolescent Psychiatry and Mental Health*, *7*(1), 29. https://doi.org/10.1186/1753-2000-7-29

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

Tapola, V., Lappalainen, R., & Wahlström, J. (2010). Brief intervention for deliberate self harm: An exploratory study. *Suicidology Online*, *1*, 95–108.

Weijters, B., Baumgartner, H., & Schillewaert, N. (2013). Reversed item bias: An integrative model. *Psychological Methods*, *18*(3), 320–334. https://doi.org/10.1037/a0032121

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample size requirements for structural equation models. *Educational and Psychological Measurement*, *73*(6), 913–934. https://doi.org/10.1177/0013164413495237

Zimmermann, P., & Iwanski, A. (2014). Emotion regulation from early adolescence to emerging adulthood and middle adulthood: Age differences, gender differences, and emotion-specific developmental variations. *International Journal of Behavioral Development*, *38*(2), 182–194. https://doi.org/10.1177/0165025413515405