The comprehensive inventory of thriving: a systematic review of published validation studies and a replication study

Angela Sorgente[1] · Michela Zambelli[1] · Semira Tagliabue[2] · Margherita Lanz[1]

Abstract

In this study we sought to collect evidence regarding the validity of the Comprehensive Inventory of Thriving (CIT), systematically reviewing studies that tested its psychometric properties (Study 1) and trying to replicate validity evidence collected across previous validation studies (Study 2). We found five studies that tested the validity of CIT scores through the collection of different kinds of evidence (score structure validity, convergent validity, discriminant validity, criterion-related validity, incremental validity, internal consistency, test-retest reliability). Results were often inconsistent across studies (especially for the score structure validity evidence). Using a sample of 483 Italian participants (63.0% female; aged 18–71 years), we replicated the tests performed in the previous validation studies. Findings suggest that the best fitting model is the one that (1) adds the overarching latent construct of thriving, which can be measured using the total scale score; and (2) merges the Skills and Flow factors in just one factor, named "Skills for Flow". At the same time, the different kinds of validity evidence collected both in previous validation studies and in the current replication study indicate high overlap among thriving sub-dimensions and poor validity evidence. We concluded that the CIT in its present form is not an adequate instrument to assess thriving, thus mono-dimensional scales (e.g. Brief Inventory of Thriving) should be currently preferred. Suggestions to develop a multi-dimensional scale measuring thriving (both using a theory-driven approach or a data-drive approach) are discussed.

Keywords Comprehensive inventory of thriving · Systematic review · Validation · Contemporary view of validity · Replication

Introduction

Well-being plays a crucial role in humans' life. It is considered one of the main predictors of health and quality of life (Cooke et al., 2016). Originally, well-being was considered to be absence of illness, that is "ill-being" (Joseph & Wood, 2010; Su et al., 2014). Subsequently, its positive connotation was also recognized, emphasizing the role of well-being in contributing to mental, physical and social functioning, beyond the absence of disease and distress (Seligman & Csikszentmihalyi, 2014). This dual nature of well-being is reflected in two approaches to well-being, namely the hedonic and the eudaimonic approaches to well-being. The *hedonic* approach deals with the experience of happiness and pleasure (Huta & Ryan, 2010) and has been linked to the so called *subjective well-being* (Keyes et al., 2002). In contrast, the *eudaimonic* approach evokes *psychological well-being*, the perception of engagement with existential challenges of life (Keyes et al., 2002). Subjective and psychological well-being have long been considered to be independent constructs, and consequently measured with different scales. Over time, the need to account for both dimensions in order to assure psychological stability and full realization of human potential has been increasingly recognized (Gabardo-Martins & Ferreira, 2018), and some authors attempted to blend together these two approaches to well-being (e.g., Kern et al., 2015; Seligman, 2018). This resulted in an urgent need to set up a comprehensive measure of well-being useful for both the scientific community and politicians (Stiglitz et al., 2009). Su et al. (2014) addressed this need by creating a new measure of well-being (the Comprehensive Inventory of Thriving; CIT), which bridges the gap between the *hedonic* and *eudaimonic* approaches by integrating all the well-being sub-dimensions within a positive and holistic approach to health.

✉ Angela Sorgente
angela.sorgente@unicatt.it

[1] Department of Pyschology, Università Cattolica del Sacro Cuore, Milan, Italy

[2] Department of Pyschology, Università Cattolica del Sacro Cuore, Brescia, Italy

Comprehensive Inventory of Thriving

According to Su et al. (2014), the term *thriving* indicates "the state of positive functioning at its fullest range—mentally, physically, and socially" (p. 256). In their theoretical framework, thriving corresponds to psychological well-being, while subjective well-being is conceived to be a sub-dimension of thriving. The multifaceted face of psychological well-being is depicted by the identification of seven core dimensions (Table 1), assessing 18 facets of positive functioning. These facets are defined and operationalized in 18 different unidimensional sub-scales of thriving, each consisting of three items, for a total of 54 items for the entire CIT scale.

Whether this model organizing thriving into 18 sub-dimensions is valid or not is still a matter of debate. To the best of our knowledge, few studies have examined the validity of CIT scores. Moreover, exhaustive conclusions about CIT scores' validity are hard to reach for two main reasons. First, while the original validation study (Su et al., 2014) proposed the 18-factor structure, successive validation studies (e.g., Hausler et al., 2017; Wiese et al., 2018) proposed alternative factorial structures. Second, there is not much evidence for some kinds of validity evidence (e.g., criterion related validity), as the validation studies subsequent to the original one (Su et al., 2014) assessed only some aspects of CIT scores' validity; for example, the recent study by Wiese

et al. (2018) examined only the cross-cultural measurement invariance of the scale, without assessing other aspects of validity, such as convergent, discriminant and criterion related validity.

The Current Study

The current study's aim is to enrich and broaden the discussion on the validity of a complex multi-dimensional measure such as the CIT, following the guidelines of the contemporary view of validity (Hubley & Zumbo, 2011; Zumbo, 2005). Two studies were conducted to meet this aim. In the first study, we performed a systematic review of the validation studies of the CIT (Study 1) to arrive at a comprehensive overview about collected evidence regarding CIT scores' validity. Then, we performed a replication study in which we tested all these kinds of validity evidence using a new sample to verify the replicability of those findings (Study 2). Indeed, performing a replication study consists of "repeat the study and see whether the results of the original and replication studies agree" (Hedges & Schauer, 2019; p. 543). The importance of the replication studies has been recently emphasized as replication is considered the best antidote to the "growing concerns over the potential unreliability of reported results in psychology" (Anderson & Maxwell, 2016; p. 1).

Study 1: Systematic Review of Comprehensive Inventory of Thriving validation studies

Method

In order to identify all the previous validation studies of the CIT scores, we conducted a systematic review of the literature, following the guidelines suggested by Siddaway et al. (2019). We used three electronic databases: Scopus, PsycINFO, and Google Scholar. In particular, we complemented the search done in Scopus and PsycINFO using Google Scholar to reach also types of scientific literature not available in the other two sources, such as syllabi and conference proceedings (Peralta-Pizza et al., 2019). The following syntax was searched in the title field of the three databases: "Comprehensive Inventory of Thriving"; no other restrictions were imposed. The search was performed in January 2020. Inclusion criteria were the following: (1) the study had to collect at least one kind of validity evidence, among those suggested by the contemporary view of validity (Hubley & Zumbo, 2011): score structure, generalizability, reliability, criterion-related, convergent/discriminant, known group, or content evidence; (2) the study had to be carried out on the entire scale of CIT (54 items) and not on a sub-sample of items.

The studies obtained from the database were screened independently by two researchers to determine whether the

Table. 1 The Seven Core Dimension and 18 Unidimensional Facets of Thriving

Seven Core Dimensions	18 Unidimensional Facets
Relationship	Support
	Community
	Trust
	Respect
	Loneliness
	Belonging
Engagement	Flow
Mastery	Skills
	Learning
	Accomplishment
	Self-Efficacy
	Self-Worth
Autonomy	(Lack of) Control
Meaning	Meaning and Purpose
Optimism	Optimism
Subjective Well-Being	Life Satisfaction
	Positive Emotions
	Negative Emotions

reported studies met the inclusion criteria. Any differences in the publications selected by the authors were discussed until agreement was reached. In particular, the two researchers extracted the same number of records from the databases and selected the same eligible records. The only record on which they needed to discuss was the paper published by Andolfi et al. (2017) as it fully respected the first inclusion criterion (i.e., collect validity evidence of CIT's score) but it did not fully respect the second inclusion criterion (i.e. to test the entire version of CIT). In particular, this paper adopted a reduced number of items and also modified some items to adapt them to a child population. After discussing with the other authors of the paper too, the two researchers decided not to include this record in the systematic review. Once the final list of selected studies was agreed upon, information about the different kinds of collected validity evidence and the studies' results was extracted from each study.

Results

A total of 25 references were obtained from the database search (eight from Scopus, 10 from PsycINFO, and seven from Google Scholar). After the duplicates were removed,

13 articles were retrieved for eligibility assessment. Only five articles met both the eligibility criteria. The list of excluded articles includes: seven papers that adopted the CIT as a well-being measure but did not collect any information about the psychometric properties of the scale, and the Andolfi et al. (2017) study, excluded for the reasons explained above. The selection flow and reasons for exclusion of sources (see Fig. 1) are documented in a Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) diagram.

Besides the original paper that discussed the development and initial validation of the CIT (Su et al., 2014), we found four other studies that reported validity evidence about this inventory. Three of them evaluated the CIT with, respectively, German (Hausler et al., 2017), Chinese (Duan et al., 2018), and Brazilian (Gabardo-Martins & Ferreira, 2018) samples. The last selected article (Wiese et al., 2018) reported a cross-cultural study that tested the measurement invariance of CIT scores across ten countries: Argentina, Australia, China, Germany, India, Mexico, Russia, Singapore, Spain, Turkey. These five selected studies were further evaluated, and the different kinds of validity evidence collected by each study, as well as the relative results, were reported in Table 2.

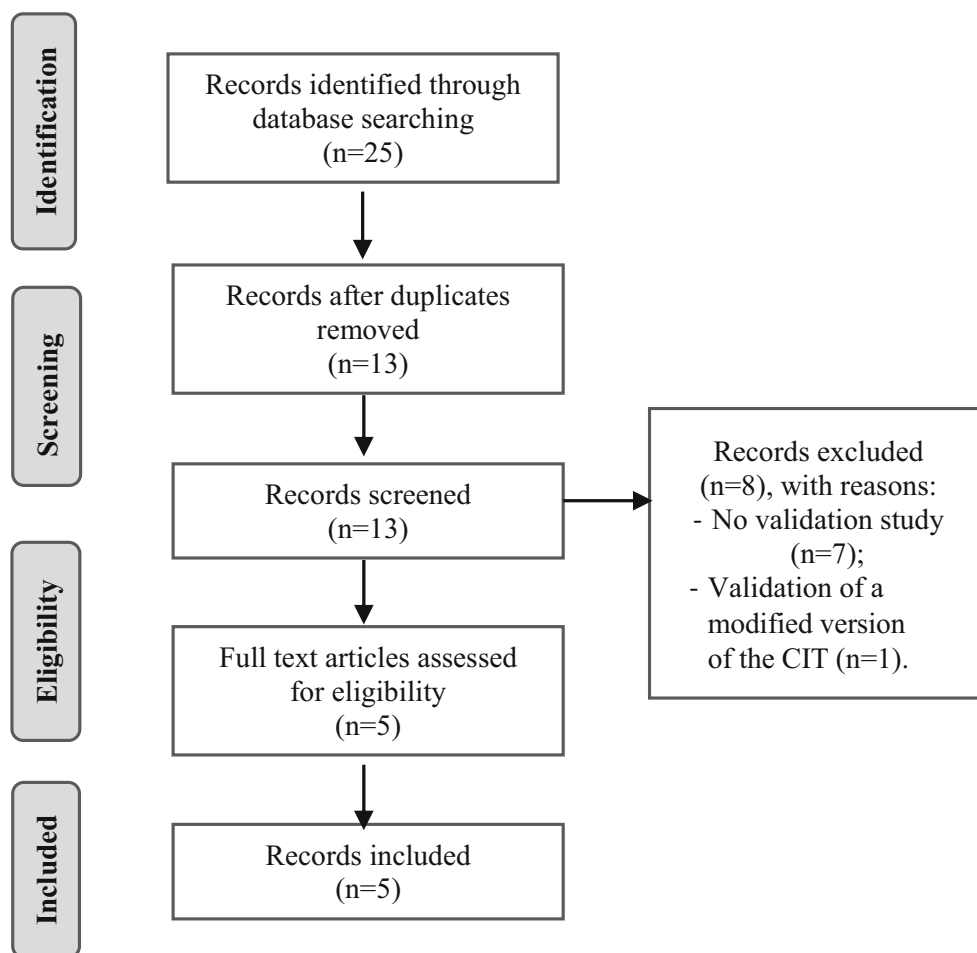

Fig. 1 PRISMA diagram of selection process

Table. 2 Kinds of validity evidence collected in each validation study and relative results

	(Su et al., 2014)	(Hausler et al., 2017)	(Duan et al., 2018)	(Gabardo-Martins & Ferreira, 2018)	(Wiese et al., 2018)
Score structure	Testing one CFA model	Testing five different CFA models	Testing two different CFA models	Testing three different CFA models	Testing four different CFA models
Generalizability					Cross-cultural measurement invariance: Partial scalar invariance for 7 of 10 countries
Convergent	Pearson correlations with other measures of well-being: r from .20 to .90	Pearson correlation with other measures of well-being: r from .16 to .80	ANOVAs comparing mean levels of well-being measures across three groups based on CIT scores (less than 25°, between 25° and 75°, more than 75° percentile): F > 15.93; $p < .001$	Spearman correlation with other measures of well-being: r from .27 to .78	
Discriminant	Pearson correlations with measures of ill-being: r from .12 to .59	Pearson correlations with measure of ill-being: r from .24 to .64	Pearson correlation with measure of ill-being: r from .34 to .63		
Criterion-related	Pearson correlations with positive and negative outcomes: r from .01 to .52		Pearson correlations with positive outcomes: r from .57 to .68		
Incremental	Regression with health outcomes: ΔR^2/Total R^2 from 29 to 92%	Regression with health outcome: ΔR^2 from 1 to 33%	Regressions with health outcomes: ΔR^2 from 3 to 48%		
Internal consistency	Cronbach's α: from .71 to .96	Cronbach's α: from .72 to .95	McDonald's ω: from .61 to .78	Cronbach's α: from .70 to .95	Cronbach's α: from .71 to .97 McDonald's ω: from .71 to .97
Test-retest	Pearson correlation: r from .57 to .83				

Note. Correlation coefficients are reported in absolute values

Across these five studies, numerous kinds of validity evidence were gathered. Some of them concerned the dimensionality of the scale, i.e., the CIT's factorial structure (score structure) and whether this structure is equivalent across subgroups of the sample (generalizability evidence). Other kinds of validity evidence consisted of proofs regarding the relationships between the CIT's sub-scales and other variables, such as measures of well-being (convergent validity), ill-being (discriminant validity) and health (criterion-validity), as well as whether the CIT scores explained health outcomes' variance more than other measures of well-being (incremental validity). Finally, the reliability of the CIT scores (internal consistency and test-retest correlation) was assessed.

Dimensionality

Based on these studies' results, the dimensionality (score structure validity) of the scale appears to be the most problematic issue with the CIT. The four validation studies which followed the original validation study by Su et al. (2014) tested and compared several alternative CFA models, thus suggesting that the CIT's factorial structure is still an open question. In particular, we identified a total of eight different factorial models across the five validation studies (Duan et al., 2018; Gabardo-Martins & Ferreira, 2018; Hausler et al., 2017; Su et al., 2014; Wiese et al., 2018). In Table 3 we specify in which validation paper we found each of these eight models. The first model was the original one proposed by Su et al. (2014) and it consists of 18 factors, one for each facet of thriving (*Model 1*). Within this model three items load on each factor; each item loads only on its respective factor and not on other factors; all factors correlate with each other, while no error covariance is included among items. The other four validation studies reported testing exactly the same model, even though (as reported in Table 3) in two cases their models resulted in a different number of degree of freedom,

Table. 3 CIT Measurement models tested across the validation studies

Model	Study	χ^2	df	p	RMSEA [90% CI]	CFI	TLI	SRMR
Model 1	Su et al. (2014)	12,757.74	1224	<.01	0.05	0.94	0.93	0.04
	Hausler et al. (2017)	7138.65	1344	<.001	0.06 [0.063, 0.066]	0.84	0.83	
	Duan et al. (2018)	$\chi^2/df = 8.59$			0.04 [0.037, 0.042]	0.90		0.05
	Gabardo-Martins and Ferreira (2018)	2144.53	1220		0.03	0.92	.91	
	Wiese et al. (2018)*	2955.78	1224		0.04	0.95	0.95	0.03
Model 2	Hausler et al. (2017)	4336.37	1342	<.001	0.05 [0.045, 0.048]	0.92	0.91	
	Gabardo-Martins and Ferreira (2018)	2699.51	1342		0.04	0.89	0.88	
Model 3	Hausler et al. (2017)	4240.68	1336	<.001	0.05 [0.044, 0.047]	0.92	0.91	
Model 4	Hausler et al. (2017)	4159.26	1331	<.001	0.04 [0.044, 0.047]	0.92	0.91	
Model 5	Hausler et al. (2017)	4493.25	1356	<.001	0.05 [0.046, 0.049]	0.91	0.91	
Model 6	Duan et al. (2018)	$\chi^2/df = 8.59$			0.05 [0.044, 0.049]	0.85		0.06
	Gabardo-Martins and Ferreira (2018)	3445.61	1356		0.04	0.83	0.82	
	Wiese et al. (2018)	12,154.18	1356		0.09	0.72	0.70	0.08
Model 7	Wiese et al. (2018)	17,146.25	1377		0.10	0.59	0.57	0.09
Model 8	Wiese et al. (2018)*	2625.13	1170		0.03	0.96	0.95	0.03

Note. df – degrees of freedom; CFI - Comparative Fit Index; RMSEA – Root Mean Square Error of Approximation; CI – Confidence Interval; SRMR – Standardized Root Mean Square Residual; TLI - Tucker-Lewis Index

*Results only referred to the American sub-sample

suggesting that some relationships were added (higher number of degree of freedom; Hausler et al., 2017) or removed/constrained (lower number of degree of freedom; Gabardo-Martins & Ferreira, 2018) in these models. Furthermore, in the cross-cultural study (Wiese et al., 2018) the original 18-factor model was tested on each single-country sample, finding a plausible solution only for seven out of 10 countries. Among these seven countries, the authors also tested the cross-cultural measurement invariance (generalizability validity), finding full metric invariance (i.e., equivalent factor loadings across countries). The second tested model (*Model 2*) consists of a hierarchical CFA model, in which the 18 sub-dimensions (first-order factors) are matched with the seven core dimensions (second-order factors). Additionally, Hausler et al. (2017) tested three more alternative hierarchical models with 18 first-order factors. *Model 3* consists of a model with 18 first-order factors, all of which load on a second-order factor called Thriving. *Model 4* is similar to Model 3, but with two second-order factors: Subjective Well-being (which three factors load on: life satisfaction, positive emotions, and negative emotions) and Psychological Well-being (which the remaining 15 factors load on). Conversely, *Model 5* is a hierarchical model composed of three levels. The first two levels, 18 first-order factors and seven second-order factors, are the same as for Model 2, while the third level consists of a third-order factor (Thriving) that all seven second-order factors load on. Unlike the other models, the sixth and the seventh models ignore the existence of the 18 sub-dimensions. In particular, *Model 6* groups the items exclusively according to the seven core dimensions of thriving (seven first-order factors), while *Model 7* estimates only one latent factor (Thriving) which all 54 items load on. Finally, Wiese et al. (2018) also tested the so-called "bi-factor model" (Liu et al., 2019). Indeed, *Model 8* specifies a general factor (Thriving) measured by all 54 items and 18 specific factors, orthogonal to the general factor, accounting for the residual variances shared by items. This model differs from Model 3, as in the bi-factor model the Thriving general factor and the 18 specific factors are not on different hierarchical levels. Wiese et al. (2018) tested this model on each single-country sample, finding admissible solutions only for four countries out of 10.

Relationships with Other Variables

The results regarding the relationships between the CIT sub-scales and other measures of well-being (convergent validity), ill-being (discriminant validity) and health (criterion-related validity) are reported in Table 2. Overall, findings suggest that the CIT sub-scales have significant relationships with most of these variables, and that these relationships are in the expected direction. Furthermore, three studies (Duan et al., 2018; Hausler et al., 2017; Su et al., 2014) verified that the CIT scores explained portions of health outcomes' variability that were not explained by other measures of well-being (incremental validity).

The main limitation we see in these validity studies is that they have not been conducted using Structural Equation Models (SEM), which allow the estimation of relationships

between latent scores that are free from measurement error (Zumbo, 2005). Moreover, we think that evidence of CIT scores' discriminant validity could be improved. Classically, discriminant validity evidence consists of testing "the strength and direction of the relationship between the measure and other variables that the measure should, theoretically, have low correlation with" (Goodwin & Leech, 2003, p.187). Indeed, a previous validation study (Duan et al., 2018) performed correlations between the CIT and measures of ill-being (e.g., the Depression Anxiety Stress Scale (DASS)), finding low negative correlations between thriving and ill-being.

However, according to recent publications and methodological reviews (e.g., Farrell, 2010; Voorhees et al., 2016; Zaiț & Bertea, 2011), when discriminant validity evidence for a multi-dimensional scale is collected within the Structural Equation Model (SEM) framework, researchers should also demonstrate that the different sub-dimensions of the scale have discriminant validity with respect to each other. In other words, once measurement error is controlled for through SEM, the relationship between each factor and its items (measuring the same thriving sub-dimension) should be stronger than the relationship this factor has with other factors (measuring different sub-dimensions).

Reliability Evidence

As reported in Table 2, CIT validation studies collected reliability evidence by testing CIT scores' internal consistency and test-retest correlation. In particular, all the reviewed studies tested the internal consistency (using Cronbach's α or McDonald's ω) of each CIT sub-scale (Table 2), while the original validation study (Su et al., 2014) was the only study that also verified the reliability of the scores by performing test-retest correlations. We see some limitations in the reliability evidence authors collected. Regarding the internal consistency estimation, it was mostly tested using the Cronbach's alpha, an index that has been recognized to be inadequate (Dunn et al., 2014) because it is based on unverified assumptions (e.g., equivalence of factor loadings). Therefore, as already done by Duan et al. (2018), Cronbach's alpha should be substituted with alternative indices, such as composite reliability (omega or ω), which is estimated within a CFA model.

Regarding the Pearson correlations performed to test the test-retest correlations, we see two limitations. First, these correlations have the same weaknesses of the correlations performed to collect other kinds of validity evidence (convergent, discriminant, etc.); that is, these correlations were performed using observed scores instead of latent scores. Second, test-retest correlations reported by Su et al. (2014) range from .57 to .83, while the minimum cut-off should be > .70 (Nunnally & Bernstein, 1994).

Despite these concerns about previous validation studies, we recognize the CIT's potential for mapping individual well-being in a comprehensive way, and we believe that its use could enrich knowledge in positive psychology. Consequently, we collected new data about the CIT (Study 2) with the aim of replicating the validity tests done in previous publications (score structure, convergent validity, discriminant validity, criterion-related validity, incremental validity, internal consistency, and test-retest correlation). We did not collect generalizability evidence (cross-cultural measurement invariance), as our sample includes only participants from Italy.

Study 2. Collecting validity evidence for the Italian version of CIT

Methods

Participants and Procedure

The sample for this study was obtained by merging two convenience samples. Participants were recruited by email and social media from institutions and organizations known by the authors. In particular, emails were sent to personal contacts of the authors and collaborators to this research project, being careful to map people living in different Italian areas (North of Italy, Central Italy, South of Italy, and islands). Any person who received the email was invited to forward the invitations to other people they know (i.e. snowball sampling). Furthermore, we also advise the current research on social media, creating in Facebook a page specifically dedicated to this project and inviting people to contact us if they were interested in participating. Those who decided to take part in the study signed an online informed consent form in accordance with the Helsinki Declaration and completed an online survey in Qualtrics, without receiving any kind of compensation for completing the survey. To avoid the same person filling the survey more than once, we used a Qualtrics option that keeps note of the respondent's IP address, thus ensuring to take the survey just one time.

The first sample included 256 participants, who were invited to fill in the CIT as well as other scales, needed to assess convergent and criterion validity. Participants in the second sample ($n = 227$) were invited to fill in only the CIT. The full sample is composed of 483 Italian participants (63.0% female) aged 18–71 ($M = 27.73$; $SD = 9.39$). Most of the participants were students (61.9%), workers (16.5%), or employed students (12.7%). The remaining participants (9.0%) belonged to several other categories (retired, looking for a job, housewife). Most participants (36.8%) had a high school diploma as the highest level of education, while others had a bachelor's degree (26.8%), a master's degree (27.9%) or a higher degree,

such as a PhD (3.9%). The remaining 4.6% of participants did not graduate from high school. To the sub-sample of 111 participants (73.4% female) aged 19–71 ($M = 28.78$; $SD = 10.55$) who gave their availability to fill in the CIT items twice, we administered again the CIT four months later in order to estimate test-retest reliability. The time span between the two assessments (four months) was the same used in the original CIT validation (Su et al., 2014).

Instruments

Comprehensive Inventory of Thriving (CIT) The Italian version of the 54-item CIT was administered. Participants were invited to indicate how much they agreed with each statement, using a 5-point Likert scale from 1= strongly disagree to 5= strongly agree. As stated in the introduction, the 54 items are grouped in 18 sub-scales (3 items per sub-scale): Support (e.g., "There are people who appreciate me as a person"), Community (e.g., "I invite my neighbors to my home"), Trust (e.g., "Most people I meet are honest"), Respect (e.g., "I am treated with the same amount of respect as others"), Loneliness (e.g., "I feel lonely"), Belonging (e.g., "I feel a sense of belonging in my community"), Flow (e.g., "In most activities I do, I feel energized"), Skills (e.g., "I use my skills a lot in my everyday life"), Learning (e.g., "I always learn something everyday"), Accomplishment (e.g., "I am on track to reach my dreams"), Self-Efficacy (e.g., "I believe that I am capable in most things"), Self-Worth (e.g., "The things I do contribute to society"), Lack of Control (e.g., "The life choices I make are not really mine"), Meaning and Purpose (e.g., "I have found a satisfactory meaning in life"), Optimism (e.g., "I have a positive outlook on life"), Life Satisfaction (e.g., "My life is going well"), Positive Emotions (e.g., "I feel happy most of the time"), Negative Emotions (e.g., "I feel negative most of the time"). Original items, instruction and response scale of the CIT (Su et al., 2014) were translated in Italian and then back translated in English, following Brislin (1986) recommendations. In order to reduce the impact that scale's translation could have had on the study's results, cognitive interviews (Ericsson & Simon, 1980; Willis, 2004) were performed with 14 Italian adults, before the definition of the final Italian pool of CIT items (see Table S1 and Table S2 of the Online Supplementary Materials for details about cognitive interviews and the Italian version of CIT items).

Psychological Well-Being Scale (PWBS) We administered the Italian version (Sirigatti et al., 2009) of the 18-item Psychological Well-being Scale developed by Ryff and Keyes (1995). This scale contains six subscales, each measured by three items: self-acceptance (e.g., "I like most aspects of my personality"), positive relationships (e.g., "I feel like I get a lot out of my friendship"), autonomy (e.g., "I am not afraid to voice my opinions, even when they are in opposition

to the opinions of most people"), environmental mastery (e.g., "I am good at juggling my time so that I can fit everything in that needs to get done"), purpose in life (e.g., "Some people wander aimlessly through life, but I am not one of them"), and personal growth (e.g., "For me, life has been a continuous process of learning, changing, and growth"). Responses are made on a 5-point Likert scale from 1 ("absolutely false") to 5 ("absolutely true"). Internal consistency was sufficient ($\omega >$ .60; Bagozzi & Yi, 1988) for each sub-scale, ranging from .64 to .70. These scales' latent scores were used to test the convergent validity of the CIT scores, as both instruments aim to assess psychological well-being sub-dimensions.

Scale of Positive and Negative Experience (SPANE) The 12-item Scale of Positive and Negative Experience (Diener et al., 2010) includes six items assessing positive feelings (e.g., joyful) and six items assessing negative feelings (e.g., sad). The Italian version of the scale (Giuntoli et al., 2017) administered to our sample was highly reliable both for the positive ($\omega =$ .87) and negative ($\omega =$ .81) factor. The SPANE latent scale's scores were used to test CIT scores' convergent validity.

Brief Symptom Inventory (BSI) The Italian version (De Leo et al., 1993) of the Brief Symptom Inventory (Derogatis & Melisaratos, 1983) was administered in order to assess the (lack of) mental health of participants. The scale is designed to evaluate psychopathological and psychological symptoms and consists of nine sub-scales: the 7-item somatization (e.g., "pain in heart or chest") subscale; the 6-item obsessive-compulsive subscale (e.g., "having to check and double-check what you do"); the 4-item interpersonal sensitivity (e.g., "feeling inferior to others") subscale; the 6-item depression (e.g., "thoughts of ending your life") subscale; the 6-item anxiety (e.g., "suddenly scared for no reason") subscale; the 5-item hostility (e.g., "getting into frequent argument") subscale; the 5-item phobic anxiety (e.g., "feeling afraid in open spaces") subscale; the 5-item paranoid ideation (e.g., "feeling that others are to blame for most of your troubles") subscale; and the 5-item psychoticism (e.g., "the idea that someone else can control your thoughts") subscale. Authors suggested that the BSI items can also be aggregated into an overall score, the General Severity Index (GSI). Composite reliability of the BSI scales ranges from .69 to .96. The BSI measure was used to collect CIT scores' criterion and incremental validity evidence.

Data Analyses

Missing data analysis, outliers deletion and descriptive statistics were performed using SPSS software. Evidence of validity was instead collected using Mplus software (version 7). *Score structure validity* was tested by performing a CFA for each of the models proposed in Table 3. Goodness of fit for

each tested model was evaluated using absolute and relative fit indices. The comparative fit index (CFI), root mean square error of approximation (RMSEA), and standardized root mean square residual (SRMR) were used to assess absolute model fit. CFIs equal to or higher than .90 and RMSEAs and SRMRs equal to or lower than .08 indicated acceptable fit. CFIs equal to or higher than .95 and RMSEAs and SRMRs equal to or lower than .05 indicated a good fit (Little, 2013). Furthermore, the AIC (Akaike's Information Criterion) and the ssBIC (sample-size adjusted Bayesian Information Criterion) were calculated for each model in order to compare models' goodness of fit, with lower AIC and ssBIC values indicating a better fit (Little, 2013).

Once the best factorial solution was selected (i.e., CFA model with best relative and absolute fit indices), the factors' scores were saved for each subject, using the SAVE DATA command in Mplus. The same procedure was used to save factors scores from other administered scales. This allowed us to test relationships among factors scores as suggested by the contemporary view of validity (Zumbo, 2005), without the necessity of performing excessively complex SEM models which were not suitable for our sample size. In particular, we gathered *convergent validity* evidence of the CIT scores by calculating Pearson correlations between the 18 factor scores of the CIT (one general factor and 17 specific factors), the six factors scores of the PWBS and the two factors scores of the SPANE.

We gathered *discriminant validity* evidence aimed at verifying that the different CIT sub-scales were well differentiated from each other (Farrell, 2010; Zaiț & Bertea, 2011). This evidence was collected by comparing the average variance extracted (AVE) by a factor from its items with the variance shared between that factor and other factors (Zaiț & Bertea, 2011). Specifically, in order to have discriminant validity: "the square root of the AVE of each construct should be much larger than the correlation of the specific construct with any of the other constructs. The value of AVE for each construct should be at least 0.50" (Zaiț & Bertea, 2011; p. 218). The AVE of a factor can be computed by averaging all of its squared factor loadings. We calculated AVE using the formula reported in Zaiț and Bertea (2011), while to calculate the shared variance between factors we squared their correlations.

Criterion-related validity of the CIT scores was evaluated by correlating each factor of the BSI with each factor of the CIT. Furthermore, each BSI factor's variance explained by the CIT factors (taken together) was estimated in a multiple regression model (Adjusted R^2). Instead, a hierarchical multiple regression was performed in order to evaluate CIT scores *incremental validity*. In other words, we assessed whether the CIT scores were able to explain portions of the mental health score's variance (Global Severity Index) that were not explained by commonly used measured of psychological well-being (PWBS; Ryff & Keyes, 1995) and subjective well-being

(SPANE; Diener et al., 2010). The GSI was set as dependent variable, the six PWBS factors scores (autonomy, environmental mastery, personal growth, purpose in life, positive relationships, self-acceptance) and the two SPANE factors scores (positive feelings, negative feelings) were entered as independent variables in the first step, while the CIT factors scores (17 specific factors and one general factor) were entered as independent variables in the second step.

Finally, CIT scores' reliability was evaluated in two ways: internal consistency, estimating composite reliability values (ω), and test-retest reliability, performing longitudinal correlations (r) in SEM models. As our longitudinal sample size (n = 111) was not sufficient to estimate the entire CIT structure for both time 1 and time 2's items in the same model, the test-retest models were run separately for each sub-dimension. In other words, for each sub-dimension we ran a model in which the latent factor was estimated both at time 1 and time 2 and the correlation between these two factors was used as a measure of test-retest reliability.

Results

Data Cleaning and Descriptive Statistics

On each item, missing data ranged from 0% (item 1) to 1.9% (item 48). These missing values were not distributed randomly: Little's MCAR test [χ^2 (2235) = 2515.52: $p < .001$]. Eight participants were removed from the sample, as their scores were univariate outliers (z absolute score was higher than 3.29; Tabachnick & Fidell, 2013) for at least two items, while 27 participants were removed because they were multivariate outliers (Mahalanobis distance based on chi-square distribution was significant for $p < .001$; Tabachnick & Fidell, 2013).

The final sample was composed of 448 participants. Descriptive statistics had been calculated for each CIT item (see Table S3 in the Online Supplementary Materials). Some of the items had a distribution that violated the assumption of normality (skewness and/or kurtosis higher than |1|). In the following analyses, missing data were handled using the full information maximum likelihood (FIML) method, while the robust maximum likelihood (MLR) estimator was used to deal with data distribution.

Score Structure Evidence

Eight different CFA models were tested for CIT in previous validation studies (see Table 3), but we ran a total of nine models on our sample as we tested two different versions of Model 8 (bi-factor model). In particular, from the number of degrees of freedom (see Table 3) of the bi-factor model proposed by Wiese et al. (2018) we inferred that they tested this model requiring correlations among single dimensions or specific factors, but, as more commonly the bi-factor model is

performed without including these correlations (see for example Sayans-Jiménez et al., 2017), we ran two versions of Model 8: Model 8a (without correlations among specific factors) and Model 8b (with correlations among specific factors). Results are reported in Table 4.

Model 8b had the best fit, followed by Model 1. Model 8b consists of a bi-factor model in which specific factors are correlated with each other. Model 1 instead coincided with the model proposed in the original CIT validation and consists of 18 factors (one for each thriving sub-dimension) correlated with each other. The problem with these two models is that they both present the following warning message in their output: "The latent variable covariance matrix (psi) is not positive definite. This could indicate a negative variance/residual variance for a latent variable, a correlation greater than or equal to one between two latent variables, or a linear dependency among more than two latent variables."

The same problem was also identified by Wiese et al. (2018): "Finally, the 18-factor model produced acceptable goodness-of-fit statistics and had admissible solutions for all but three countries (Argentina, Mexico, and China). Specifically, these latter three countries had a covariance matrix that was not positive definite, which may indicate a linear dependency between two or more factors in these samples" (p. 137). While the decision of Wiese et al. (2018) was to drop these three countries from the analysis, we think that the warning message should be used to better understand thriving's sub-dimensions and their relationships. Since the warning message suggested that the problem was related to the Skills latent factor, we checked its relationships with other factors and we found a very high correlation between Skills and Flow latent factors ($r = .91$; $p < .001$), showing that these two latent factors are linearly dependent (as suggested by the warning message) and they are measuring the same construct. We decided to run two new models (respectively "Model 8b – modified" and "Model 1 – modified"; see Table 4) in which the three items of Skills and the three items of Flow loaded on the same latent factor, which we named "Skills for Flow", as the literature suggests that having a sufficient level of skill in a specific task is a pre-requisite for experiencing flow while that task is being performed (Fullagar et al., 2013). The "Model 8b – modified" consists of one general factor (thriving) and 17 specific factors correlated with each other, while "Model 1 – modified" consists only of the 17 correlated factors without the general factor. In both the modified models, the warning message disappeared, confirming that the relation between Skills and Flow latent factors was the problematic one. Among the two modified models, "Model 8b – modified" presented the admissible solution with the best relative and absolute fit indices, so we selected this factorial solution as the final model of CIT. We reported items' factor loadings and correlations among latent factors for the "Model 8b – modified" respectively in Table S4 and Table S5 of the Online Supplementary Materials. Furthermore, we saved for each participant the latent factors' scores in order to use them to collect the following kinds of validity evidence.

Convergent Validity

Correlations between the 18 factors scores of the CIT (one general and 17 specific factors), the six factors scores of the PWBS and the two factors scores of the SPANE are reported in Table 5. These correlations were performed on the sample of participants ($n = 242$) who also filled in scales other than the CIT. Results partially support the validity of CIT scores; from one side, detected relationships are in the expected direction, but, on the other side, they sometimes were non-significant. The general factor "Thriving" has a significant positive relationship with all the sub-dimensions of the PWBS (psychological well-being). Instead, in relation to the emotional component of the subjective well-being (SPANE), the "Thriving" factor has a significant positive relationship with the "Positive Feeling" factor, but a non-significant negative relationship with the "Negative Feeling" factor ($r = -.13$; $p = .058$). The different sub-scales of the CIT had always at least a non-significant relationship with the sub-dimensions of psychological and subjective well-being. The "Loneliness" factor is the only CIT sub-scale that has a significant relationship with all the dimensions of the PWBS and SPANE (see Table 5).

Discriminant Validity

Evidence of discriminant validity is reported in Table 6. On the diagonal of the table, the AVE value is reported for each CIT sub-scale. This value represents the amount of variance that each latent factor explains in relation to the observed items that belong to that latent factor. As stated by Zaiț and Bertea (2011), a sufficient level of discriminant validity requires that each factor's AVE value is both higher than .50 and much larger than the amount of variance that the factor shares with other factors (see squared correlations under the diagonal). Only three CIT sub-scales (Learning, Negative emotions and Self-worth) satisfied this requirement. This means that other latent factors do not sufficiently explain responses that participants give to their items (e.g., "Support" factor has an AVE lower than .50) and/or that CIT latent factors share more variance with other latent factors than with their own items. For example, the "Life Satisfaction" factor shares more variance with the "Optimism" factor (squared correlation = .63) than with items measuring life satisfaction (AVE = .59).

—

Table. 4 Score structure validity evidence for the Italian CIT (N = 445)

Model	AIC	ssBIC	χ^2	df	RMSEA [90% CI]	CFI	TLI	SRMR
Model 1*	46,449.29	46,740.51	2172.69	1224	.04 [.039 .045]	.90	.88	.05
18-factor model								
Model 1 – modified	46,480.19	46,755.69	2228.86	1241	.04 [.039 .045]	.89	.88	.05
17-factor model								
Model 2	47,055.97	47,238.10	2946.32	1342	.05 [.049 .054]	.83	.81	.07
Hierarchical model (7 second-order factor and 18 first-order factor)								
Model 3	47,268.04	47,434.45	3161.96	1359	.05 [.052 .057]	.81	.79	.07
Hierarchical model (1 second-order factor and 18 first-order factor)								
Model 4	47,224.72	47,392.06	3121.51	1358	.05 [.052 .057]	.81	.80	.07
Hierarchical model (2 second-order factor and 18 first-order factor)								
Model 5	47,147.13	47,316.31	3049.76	1356	.05 [.050 .055]	.82	.81	.07
Hierarchical model (1 third-order factor, 7 second-order factor and 18 first-order factor)								
Model 6	48,479.65	48,648.84	4263.72	1356	.07 [.067 .072]	.69	.67	.08
7-factor model								
Model 7	49,693.24	49,843.01	5386.29	1377	.08 [.079 .083]	.57	.55	.09
1-factor model								
Model 8a	47,200.03	47,399.72	3042.99	1323	.05 [.052 .057]	.81	.80	.07
Bi-factor model (1 general factor and 18 specific factors) without correlations								
Model 8b*	46,263.52	46,604.67	1926.54	1170	.04 [.035 .041]	.92	.90	.04
Bi-factor model (1 general factor and 18 specific factors) with correlations								
Model 8b – modified	46,285.94	46,611.37	1981.94	1187	.04 [.036 .042]	.91	.90	.04
Bi-factor model (1 general factor and 17 specific factors) with correlations								

Note. AIC – Akaike's Information Criterion; ssBIC – sample-size adjusted Bayesian Information Criteria; *df* – degrees of freedom; RMSEA – Root Mean Square Error of Approximation; CI – Confidence Interval; CFI - Comparative Fit Index; TLI - Tucker-Lewis Index; SRMR – Standardized Root Mean Square Residual.

*The model is estimated but a warning message is reported in the output.

Table. 5 Pearson correlations between CIT and PWBS and SPANE factors scores

	PWBS						SPANE	
	AU	EM	PG	PL	PR	SA	PF	NF
Thriving	.22**	.16*	.56***	.37***	.38***	.34***	.23***	−.12
Support	.03	.22***	.19***	.26***	.53***	.28***	.28***	−.22***
Community	−.04	.13*	−.03	.23***	.05	.10	.24***	−.16*
Trust	−.03	.21***	.08	.23***	.35***	.34***	.40***	−.33***
Respect	.11	.27***	.08	.28***	.30***	.42***	.40***	−.35***
Loneliness	−.26***	−.49***	−.13*	−.36***	−.36***	−.51***	−.54***	.58***
Belonging	−.12	.14*	.01	.20**	.19**	.23***	.26***	−.21**
Skills for Flow	.12	.49***	.07	.69***	.19**	.61***	.51***	−.45***
Learning	.01	.11	.14*	.34***	.13*	.20**	.13*	−.03
Accomplishment	.02	.45***	.08	.62***	.24***	.62***	.46***	−.44***
Self-efficacy	.23***	.47***	.08	.61***	.14*	.64***	.53***	−.52***
Self-worth	−.07	.25***	.08	.47***	.14*	.33***	.30***	−.20**
Control	.22**	.40***	.13	.28***	.21**	.34***	.29***	−.29***
Meaning and Purpose	.07	.50***	.12	.66***	.29***	.65***	.57***	−.54***
Optimism	.09	.45***	.10	.53***	.24***	.70***	.66***	−.67***
Life Satisfaction	.05	.50***	.08	.59***	.29***	.69***	.59***	−.59***
Positive emotions	.09	.44***	.11	.50***	.28***	.64***	.70***	−.67***
Negative emotions	−.20**	−.48***	−.10	−.45***	−.20**	−.57***	−.66***	.66***

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

PWBS = Psychological Well-being Scale; SPANE = Scale of Positive and Negative Experience, AU = autonomy, EM = environmental mastery, PG = personal growth, PL = purpose in life, PR = positive relationships, SA = self-acceptance, PF = positive feelings, NF = negative feelings

Table. 6 Average variance explained (AVE) by each CIT sub-scale and shared variance among sub-scales (N = 445)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1. Support	**.38**																
2. Community	0	**.31**															
3. Trust	.15	.16	**.42**														
4. Respect	.26	0	.45	**.38**													
5. Loneliness	.22	0	.03	.16	**.47**												
6. Belonging	.05	.41	.24	.07	.04	**.38**											
7. Skills for Flow	.09	.23	.14	.11	.07	.14	**.39**										
8. Learning	.04	.07	.01	.005	.004	.04	.21	**.51**									
9. Accomplishment	.16	.03	.11	.11	.08	.09	.49	.07	**.64**								
10. Self-efficacy	.02	.08	.06	.12	.08	.03	.55	.09	.26	**.30**							
11. Self-worth	.08	.15	.11	.07	.002	.16	.33	.09	.22	.08	**.62**						
12. Control	.03	.004	.01	.04	.25	0	.02	.005	.09	.01	.002	**.48**					
13. Meaning and Purpose	.16	.08	.10	.08	.20	.13	.51	.09	.63	.27	.30	.05	**.48**				
14. Optimism	.04	.06	.16	.10	.22	.15	.36	.02	.38	.45	.07	.03	.46	**.42**			
15. Life Satisfaction	.17	.07	.15	.13	.28	.19	.46	.04	.72	.24	.15	.10	.74	.63	**.59**		
16. Positive emotions	.13	.05	.19	.17	.28	.11	.32	.02	.33	.30	.11	.04	.52	.75	.63	**.69**	
17. Negative emotions	.02	.01	.03	.07	.56	.03	.14	.01	.16	.18	.01	.18	.22	.46	.31	.59	**.68**

Note. On the diagonal, in bold, AVE values (i.e., average variance explained by each factor) are reported, while under the diagonal squared correlation between couple of factors are reported

Criterion-Related Validity

In order to assess whether CIT scores were able to predict the scores of a current criterion measure (BSI), we estimated the percentage of variance (Adjusted R^2) that CIT factors scores (taken together) explain of each BSI factor score. We found that the individual level of thriving, measured through CIT factors, explains from 19.1% (for the "Phobic anxiety" sub-scale) to 55.4% (for the "Depression" sub-scale) of individual mental health, thus confirming the criterion-related validity of the CIT scores. Further evidence about this kind of validity is presented in Table S6 of the Online Supplementary Materials, where each BSI factor's adjusted R^2 and correlations between each CIT factors and BSI factors are reported. Interestingly, as for the convergent validity evidence, the relationships between variables were always in the expected direction but they were not always significant. Particularly, we had some CIT sub-dimensions (e.g., Learning, Community) having non-significant relationships with most of the criterion-related variables.

Incremental Validity

Results of the hierarchical regression suggested that the subscales of PWBS and SPANE were significant predictors of the GSI [$F (8, 231) = 26.23; p < .001$] explaining the 48.1% of the dependent variable's variance ($R^2 = .48$). When CIT scores were added as independent variables, the regression model significantly [$F (26, 213) = 11.08; p < .001$] explained 57.5% ($R^2 = .57$) of the GSI's variance. The significant increment in the variance explained by the regression model (change of $R^2 = .09; p = .001$) suggests that the CIT scores have incremental validity when compared with common measures of psychological and subjective well-being. Note that this evaluation refers to the 18 CIT factors (17 specific factors and one general factor) taken together and not individually. For details about the regression relationships that each factor had with the GSI, see Table S7 of the Online Supplementary Materials.

Reliability Evidence

As reported in Table 7, all the CIT sub-scales had sufficient internal consistency ($\omega > .60$; Bagozzi & Yi, 1988), except for the "Community" and "Self-efficacy" factors. Furthermore, test-retest reliability was sufficient ($r > .70$; Nunnally & Bernstein, 1994) for all the sub-scales, except for "Control", "Respect", and "Support" factors. Thus, also for this kind of evidence, results are not promising for all the CIT sub-dimensions.

Table. 7 Reliability evidence for the CIT scores

Latent Factor	Composite reliability	Test-retest correlation
Thriving	.73	
Support	.65	.68***
Community	.56	.88***
Trust	.69	.82***
Respect	.65	.57***
Loneliness	.73	.94***
Belonging	.64	.76***
Skills for Flow	.79	.77***
Learning	.62	.75***
Accomplishment	.84	.80***
Self-efficacy	.57	.92***
Self-worth	.77	.76***
Control	.73	.68***
Meaning and Purpose	.72	.87***
Optimism	.68	.998***
Life satisfaction	.81	.78***
Positive emotions	.87	.85***
Negative emotions	.86	.88***

Note. Test-retest correlation for the thriving factor was not calculated as it was not possible to estimate in the same model two 54-item factors on a sample of 111 cases; *** $p < .001$

Discussion

In the current paper, we aimed to evaluate whether CIT scores yielded valid measures of thriving. To reach this goal, two different studies were carried out. The first study (Study 1) consisted of a systematic review that discussed five different articles which previously tested the psychometric characteristics of the CIT. Study 2 tested the different kinds of validity evidence (score structure, convergent, discriminant, criterion-related, incremental, internal consistency, test-retest) that were previously tested across the validation studies of CIT with a new sample. Here the results of the previous five validation studies and the current replication study are discussed to conclude if the results about CIT validity are credible and generalizable or if they are "context specific" (Aguilar, 2020).

Dimensionality

The *score structure validity* is the only kind of validity evidence (other than reliability) that has been tested in all the five validation studies we collected in Study 1. Results obtained across these studies as well as our replication study (Study 2) denounce the lack of score structure validity evidence for the CIT. In particular, the original 18 factor model has been rarely replicated. As reported in Table 3, all the validation studies

(Duan et al., 2018; Gabardo-Martins & Ferreira, 2018; Hausler et al., 2017; Wiese et al., 2018) run this model, but - to obtain a satisfactory fit - authors needed to slightly modify the model, resulting in a higher (Hausler et al., 2017) or lower (Gabardo-Martins & Ferreira, 2018) number of degree of freedom than in the original model. When instead the original model was fully respected (same degree of freedom; Wiese et al., 2018) the fit indices indicated that the model did not represent well all the data (it was a plausible solution only for seven out of 10 countries included in Wiese et al. (2018) study).

These unsatisfactory results have generated the proposal of seven factorial models alternative to the original one across the different validation studies (see Table 3). We replicated both the original model and these alternative models, never finding satisfactory fit indices. Furthermore, concerns about the score structure validity of the CIT arose also from the linear dependence among sub-dimensions of the CIT that both we (Study 2) and Wiese et al. (2018) found testing the CIT factorial models. This linear dependence indicates that the CIT sub-dimensions are someway overlapped with each other, as denounced also by other results we collected. First, both in our sample and in Wiese et al.'s (2018) study, the bi-factor model – the model that best takes into account the overlap among the 18 factors, by providing both a general factor (which explains all the common variance shared by all the specific factors) and correlations between each pair of specific factors (that account for the variance shared only by each couple of factors) – is the one which best describes the data. Such model was not tested in other CIT validation studies (Duan et al., 2018; Gabardo-Martins & Ferreira, 2018; Hausler et al., 2017; Su et al., 2014). Second, we found that, at least in our sample (Study 2), two factors (Skills and Flow) that should be theoretically distinct are actually measuring the same construct (their correlation approached 1). Finally, we found that CIT scores (except for Learning, Negative Emotions and Self-Worth factors) lacked *discriminant validity*, i.e., CIT items do not discriminate well among the different thriving facets they should measure. For example, the "Life Satisfaction" latent factor shares more variance with the "Optimism" latent factor (squared correlation = .63) than with items measuring life satisfaction (AVE = .59).

Relationship with Other Variables

In this section we discuss results concerning the relationship that the CIT scores have with scales measuring the same construct (*convergent validity evidence*), with scales measuring criterion constructs (*criterion-related validity evidence*) as well as results concerning the ability of the CIT scores to explain portions of criterion constructs' variability that were not explained by other measures of well-being (*incremental validity evidence*). These three kinds of validity evidence were already tested respectively in four (Duan et al., 2018;

Gabardo-Martins & Ferreira, 2018; Hausler et al., 2017; Su et al., 2014), two (Duan et al., 2018; Su et al., 2014), and three (Duan et al., 2018; Hausler et al., 2017; Su et al., 2014) validation studies included in Study 1. These kinds of validity evidence found the highest agreement among studies, as the reported relationships were always in the expected direction. At the same time, these results are not fully trustable for the following reasons. First, these relationships were not always significant (e.g., Su et al., 2014). Second, some studies (e.g., Duan et al., 2018) tested these kinds of validity evidence using the entire scale score and not each specific sub-scale score. Similarly, in our replication study (Study 2), we found that expected results about the relationships between the CIT factors and other measures are confirmed only when the CIT factors are evaluated altogether. Instead, if we look for evidence about each specific factor, we must conclude that some sub-dimensions of CIT (such as "Learning" and "Community") lack both convergent and criterion-related validity, as they are not significantly related to measures of well-being and health. We drew similar conclusions for the incremental validity evidence as we found that the 18 factors (one general factor and 17 specific factors) of the CIT successfully explained a portion of the health outcomes that other measures of well-being (PWBS and SPANE) did not explain only if the 18 CIT factors were considered together. On the contrary, when their incremental validity was evaluated by checking against the single factor (see Table S7 of the Online Supplementary Material), we found that most of the CIT dimensions did not explain variance in the dependent variable (mental health). All in all, convergent, criterion-related, and incremental validity evidence of the CIT scores is not sufficient for the multidimensional structure of the scale.

Reliability Evidence

As reported in Table 2, two different ways to test the reliability evidence of the CIT scores were adopted across validation studies. All the five studies included in Study 1 assessed the internal consistency of the scale, finding sufficient level of reliability. When we tried to replicate these results (Study 2), we found that Self-efficacy and Community factors had lower reliability levels than the minimum cut-off (ω = .60; Bagozzi & Yi, 1988). This discrepancy from previous studies could depend on the different method used to collect internal consistency (alpha vs. omega). However, as we adopted the method suggested by the contemporary view of validity (i.e. omega, as it allows to remove the measurement error and give estimations that are more precise), the internal consistency of CIT seems to be questionable.

The second way to assess CIT reliability we found in the validation papers included in Study 1 is the test-retest correlation. This kind of evidence was collected only by Su et al. (2014), who found that the CIT test-retest correlations not always reached the minimum cut-off (r = .70; Nunnally &

Bernstein, 1994). Our replication study (Study 2) confirmed this lack of reliability as we found that "Support", "Respect", and "Control" factors had test-retest correlations lower than .70. In sum, both the internal consistency and the test-retest stability of CIT seem to be questionable.

General Discussion

All in all, results obtained in the previous validation studies (Study 1) as well as in our replication study (Study 2) suggest that CIT validity evidence is not trustworthy and generalizable. We believe that this conclusion mainly depends on the complexity intrinsic to the well-being construct. Well-being is a multi-dimensional construct, where the different dimensions are (at least at some degree) overlapped each other (e.g., Kállay & Rus, 2014; Schrank et al., 2013). This overlap contrasts with the theoretical assumptions of CFA framework, where each item is expected to be explained by only one latent factor.

The conceptual overlap among well-being sub-dimensions is a problem that CIT shared with other well-being measures, such as Ryff (1989)'s Psychological Well-Being Scale (PWBS). In particular, Springer and Hauser (2006) performed a replication study to test the validity evidence of the PWBS scale as previous validation studies found unsatisfactory and contradictory results. They concluded that there is little support for the theoretically proposed multidimensionality of the scale, due to the "very high overlap among dimensions" (p. 1080). Our study, in agreement with Springer and Hauser (2006)'s study, confirms the difficulty to empirically support the theoretical multidimensionality of the well-being scales. In sum, validity evidence of CIT scores collected in the previous validation studies as well as in the current research indicates that the CIT, as already happen to other multidimensional scales assessing well-being, is not able to guarantee an empirical assessment that fully adhere to the theoretical multidimensional model authors propose.

All that considered, which are the main implications for researchers and practitioners? If the researcher or practitioner is interested in measuring thriving, we suggest adopting the Brief Inventory of Thriving (BIT) instead of the CIT. The BIT has been developed in conjunction with the CIT (Su et al., 2014) and consists of a 10-item unidimensional scale that assesses thriving (in Table S2 of the Online Supplementary Materials we have specified which are the 10 items of CIT that compose the BIT). The BIT showed good psychometric properties in the original (Su et al., 2014) and successive (Duan et al., 2016; Gabardo-Martins & Ferreira, 2018; Sorgente et al., 2019; Sorgente et al., 2020) validations. The adoption of a mono-dimensional scale to assess different facets of well-being has been suggested also in previous publications (e.g., Hills & Argyle, 2002). Well known is the Flourishing Scale, a brief 8-item scale that summaries important areas of well-being such as relationships, self-esteem, purpose, and optimism (Diener et al., 2010). Recognizing the multidimensionality of well-being, these brief scales simplify the assessment of well-being by including items that measure different well-being facets so that the latent factor on which these items converge, represents only what is shared among these facets.

The preference for the mono-dimensional scales over the multidimensional ones is evident when looking for studies that adopted respectively the CIT and the BIT. Few studies adopted the CIT scale to measure well-being in cross-sectional study (Iannello et al., 2021) or to assess the effectiveness of welfare interventions, such as the effectiveness of "best possible self" intervention with psychology undergraduate students (Heekerens & Heinitz, 2019). Furthermore, these studies have not adopted the entire CIT scale, but only a few subscales of the CIT. Instead, an extensive use of BIT is registered, both to measure well-being in cross-sectional study (e.g., Duan et al., 2018; Duan & Bu, 2019; Höfer et al., 2019; Sorgente & Lanz, 2019; Ugwu et al., 2018) and to assess the effectiveness of welfare interventions, such as a therapeutic horticulture intervention with child survivors of sexual abuse (Watkins et al., 2019), a 9-week trauma and resilience curriculum with high school students (Judge, 2018), a single-session positive cognitive intervention (Bu & Duan, 2019) as well as a character strength-based intervention with first-year university students (Duan et al., 2019), a fourteen-module video-recorded well-being intervention with undergraduate students (Singh & Bandyopadhyay, 2020), a group intervention for refugees and asylum seekers (Reebs et al., 2020), and a training to foster teachers' well-being (Rahm & Heise, 2019).

In case the researcher and/or practitioner is interested in measuring the different dimensions of thriving, we suggested not to use the CIT scale as the current study demonstrates that the validity of this scale lack of sufficient evidence and replication. We think that the measurement of the multidimensional thriving is far to be solved and we see two possible alternatives for future psychometric studies. The first alternative consists of adopting a *theory-driven approach*, like the one that bases the validation studies described in the current paper, but paying more attention to make the items as differentiated as possible, in order to reduce the overlap among dimensions. In this case, scale developers could maintain the 18 facets model of the CIT (theory-driven approach), but should (1) rewrite items in order to modify those that seem to measure more than one dimension (like the Skills and Flow factor's items, which in the current study show to measure the same construct); and (2) conduct cognitive interviews with the target population questioning potential respondents about the meaning of the items (Ericsson & Simon, 1980; Willis, 2004). Indeed, what seems theoretically distinguishable from the researchers' perspective, could be experienced by respondents in a less differentiated way. After that items have been modified

following the above mentioned criteria, researchers can perform theory-driven analysis (e.g., CFA) to confirm the expected multi-dimensionality of the construct. In case this items' modification would be not sufficient (e.g., Springer and Hauser (2006) found that the high overlap among PWBS dimensions remained even when controlling for question wording, question order, negative item-wording, item placement and item redundancy), researchers should consider the second alternative here proposed.

The second alternative consists in adopting a *data-driven approach*, such as the recently proposed Network Psychometrics (Epskamp et al., 2018), which "allows the model structure to spontaneously emerge from the relationships among indicators" (Giuntoli & Vidotto, 2021, p. 896). In this perspective, the items in our questionnaire are not a function of the latent variable (i.e., people's responses to items are not caused by their position on the latent variable), but rather there is a pattern of mutual dependencies among observed variables (items) from which the latent construct emerges (Christensen et al., 2020; Cramer et al., 2012). In other words, strong relationships between items belonging to different sub-dimensions are not problematic (e.g. overlap among sub-dimension), but these relationships build the structure that will define the latent construct (data-driven approach). As Giuntoli and Vidotto (2021) stressed, Network Psychometrics "is a particularly interesting approach to answer the question on the dimensionality of well-being, because instead of testing the alternative hypotheses of multidimensionality versus one-dimensionality of the well-being construct, we can "let the data speak by themselves."" (p. 904).

Summing up, we believe the current study has many implications for researchers as well as practitioners. On one side, researchers have to use the theory-driven (e.g., items reformulation, cognitive interviews) and data-driven (e.g., Network Psychometrics) solutions presented above in order to find a balance between theoretical well-being models which propose multi-dimensional structure and what data are suggesting (i.e. overlap between dimensions).

On the other side, practitioners have to take into consideration the overlap between well-being dimensions in their clinical and treatment practice. In particular, from the clinical point of view they cannot think to work on a client/patient's single well-being dimension as it would be not related to the others. For example, Margola et al. (2019) study found that three sub-dimensions of the family well-being (communication, cohesion, lack of conflict) were perceived as a mono-dimensional factor (relationship quality) at family level. These results support clinicians who adopt a holistic approach in their clinical practices (e.g. Sultanoff, 1997), encouraging the integration of the physical, mental, emotional, and spiritual aspects of the human condition in treatment.

This approach has also consequences in the evaluation of treatment effectiveness. The well-being of clients/patients should be evaluated pre- and post-treatment using comprehensive scales which are able to assess diverse well-being dimensions at once. We suggested to use the Brief Inventory of Thriving because (1) it assesses numerous dimensions of well-being (life satisfaction, optimism, positive emotion, belonging, support, self-worth, flow, accomplishment, meaning and purpose) using only 10 items; (2) its test-retest reliability has been demonstrated in different studies (e.g., Sorgente et al., 2018; Su et al., 2014); (3) it is commonly adopted to evaluate the efficacy of treatments (e.g., Bu & Duan, 2019; Duan et al., 2019; Judge, 2018; Rahm & Heise, 2019; Reebs et al., 2020; Singh & Bandyopadhyay, 2020; Watkins et al., 2019).

Limitations and Strenghts

The main limitation we identified in our study is the sample we adopted. It is both a convenience sample and a sample from just one country (Italy). This could lead someone to argue that our results are not generalizable. We think that our study can be considered a piece of evidence that, together with other validation studies of the CIT (which in 80% of the cases are based on just one country; Su et al., 2014; Duan et al., 2018; Gabardo-Martins & Ferreira, 2018; Hausler et al., 2017), can help with drawing conclusions about the validity of this scale. Furthermore, we argue that our study has some strength compared to the other validation studies. First, our study is the only one in which cognitive interviews aiming to verify that the target population has a clear understanding of CIT (translated) items have been conducted. Second, other than the original study (Su et al., 2014), our study is the only one examining all the different kinds of validity evidence (score structure, convergent, discriminant, criterion, incremental, internal consistency, test-retest). Third, our study is the only one examining all the measurement models proposed across different studies using the same sample. Finally, our study is fully based on current guidelines regarding the contemporary view of validity.

Conclusion

The current paper provides a general overview of the CIT's psychometric properties, as it both reviewed the results of previous validation studies and replicated the validity evidence collected in those studies. Findings suggest that the CIT scale has numerous weaknesses, as each kind of the validity evidence here investigated show results that are not satisfactory and/or not consistent across studies. The idea that thriving is composed of 18 facets that are distinct from each other was not confirmed, due to the strong overlap found among CIT sub-dimensions. Currently, we suggest to

researchers and practitioners to assess thriving using the mono-dimensional BIT while working to enhance the multi-dimensional measurement of thriving by improving the differentiation of items when using a theory-driven approach or by adopting a data-driven approach.

Declarations

References

Aguilar, M. S. (2020). Replication studies in mathematics education: What kind of questions would be productive to explore?. *International Journal of Science and Mathematics Education, 18*, 37–50. https://doi.org/10.1007/s10763-020-10069-7.

Anderson, S. F., & Maxwell, S. E. (2016). There's more than one way to conduct a replication study: Beyond statistical significance. *Psychological Methods, 21*(1), 1–12. https://doi.org/10.1037/met0000051.

Andolfi, V. R., Tay, L., Confalonieri, E., & Traficante, D. (2017). Assessing well-being in children: Italian adaptation of the comprehensive inventory of thriving for children (CIT-CHILD). *TPM: Testing, Psychometrics, Methodology in Applied Psychology, 24*(1), 127–145. https://doi.org/10.4473/TPM24.1.8.

Bagozzi, R. P., & Yi, Y. (1988). On the evaluation of structural equation models. *Journal of the Academy of Marketing Science, 16*(1), 74–94. https://doi.org/10.1177/009207038801600107.

Bu, H., & Duan, W. (2019). A single-session positive cognitive intervention on first-year students' mental health: Short-term effectiveness and the mediating role of strengths knowledge. *Journal of American College Health, 67*(6), 515–522. https://doi.org/10.1080/07448481.2018.1497639.

Christensen, A. P., Golino, H., & Silvia, P. J. (2020). A psychometric network perspective on the validity and validation of personality trait questionnaires. *European Journal of Personality, 34*(6), 1095–1108. https://doi.org/10.1002/per.2265.

Cooke, P. J., Melchert, T. P., & Connor, K. (2016). Measuring well-being: A review of instruments. *The Counseling Psychologist, 44*(5), 730–757. https://doi.org/10.1177/0011000016633507.

Cramer, A. O., Van der Sluis, S., Noordhof, A., Wichers, M., Geschwind, N., Aggen, S. H., et al. (2012). Dimensions of normal personality as networks in search of equilibrium: You can't like parties if you don't like people. *European Journal of Personality, 26*(4), 414–431. https://doi.org/10.1002/per.1866.

De Leo, D., Frisoni, G. B., Rozzini, R., & Trabucchi, M. (1993). Italian community norms for the brief symptom inventory in the elderly. *British Journal of Clinical Psychology, 32*(2), 209–213. https://doi.org/10.1111/j.2044-8260.1993.tb01045.

Derogatis, L. R., & Melisaratos, N. (1983). The brief symptom inventory: An introductory report. *Psychological Medicine, 13*(3), 595–605. https://doi.org/10.1017/S0033291700048017.

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

Duan, W., & Bu, H. (2019). Randomized trial investigating of a single-session character-strength-based cognitive intervention on Freshman's adaptability. *Research on Social Work Practice, 29*(1), 82–92. https://doi.org/10.1177/1049731517699525.

Duan, W., Bu, H., Zhao, J., & Guo, X. (2019). Examining the mediating roles of strengths knowledge and strengths use in a 1-year single-session character strength-based cognitive intervention. *Journal of Happiness Studies, 20*(6), 1673–1688. https://doi.org/10.1007/s10902-018-0014-z.

Duan, W., Fei, Y., Zhao, J., & Guo, X. (2018). Incremental validity of the comprehensive inventory of thriving in predicting self-reporting mental and physical health among community populations. *Journal of Health Psychology., 25*, 1366–1373. https://doi.org/10.1177/1359105318755265.

Duan, W., Guan, Y., & Gan, F. (2016). Brief inventory of thriving: A comprehensive measurement of wellbeing. *Chinese Sociological Dialogue, 1*(1), 15–31. https://doi.org/10.1177/2397200916665230.

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology, 105*(3), 399–412. https://doi.org/10.1111/bjop.12046.

Epskamp, S., Maris, G., Waldrop, L. J., & Borsboom, D. (2018). Network psychometrics. In P. Irwing, D. Hughes, & T. Booth (Eds.), *The Wiley handbook of psychometric testing, 2 volume set: A multidisciplinary reference on survey, scale and test development*. Wiley. https://doi.org/10.1002/9781118489772.ch30.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review, 87*(3), 215–251. https://doi.org/10.1037/0033-295X.87.3.215.

Farrell, A. M. (2010). Insufficient discriminant validity: A comment on Bove, Pervan, Beatty, and Shiu (2009). *Journal of Business Research, 63*(3), 324–327. https://doi.org/10.1016/j.jbusres.2009.05.003.

Fullagar, C. J., Knight, P. A., & Sovern, H. S. (2013). Challenge/skill balance, flow, and performance anxiety. *Applied Psychology, 62*(2), 236–259. https://doi.org/10.1111/j.1464-0597.2012.00494.x.

Gabardo-Martins, L. M. D., & Ferreira, M. C. (2018). Psychometric properties of the inventory of thriving: Brief and comprehensive versions. *Trends in Psychology, 26*(3), 1573–1587. https://doi.org/10.9788/TP2018.3-15En.

Goodwin, L. D., & Leech, N. L. (2003). The meaning of validity in the new standards for educational and psychological testing:

Implications for measurement courses. *Measurement and Evaluation in Counseling and Development, 36*(3), 181–191. https://doi.org/10.1080/07481756.2003.11909741.

Giuntoli, L., & Vidotto, G. (2021). Exploring Diener's multidimensional conceptualization of well-being through network psychometrics. *Psychological Reports, 124*(2), 896–919. https://doi.org/10.1177/0033294120916864.

Giuntoli, L., Ceccarini, F., Sica, C., Open, C. C.-S., & 2017, undefined. (2017). Validation of the Italian versions of the flourishing scale and of the scale of positive and negative experience. SAGE Open*, 7*(1), 1–12. doi: https://doi.org/10.1177/2158244016682293.

Hausler, M., Huber, A., Strecker, C., Brenner, M., Höge, T., & Höfer, S. (2017). Validierung eines Fragebogens zur umfassenden Operationalisierung vonWohlbefinden: Die deutsche Version des Comprehensive Inventory of Thriving (CIT) und die Kurzversion Brief Inventory of Thriving (BIT). *Diagnostica, 63*(3), 219–228. https://doi.org/10.1026/0012-1924/a000174.

Hedges, L. V., & Schauer, J. M. (2019). More than one replication study is needed for unambiguous tests of replication. *Journal of Educational and Behavioral Statistics, 44*(5), 543–570. https://doi.org/10.3102/1076998619852953.

Heekerens, J. B., & Heinitz, K. (2019). Looking forward: The effect of the best-possible-self intervention on thriving through relative intrinsic goal pursuits. *Journal of Happiness Studies, 20*(5), 1379–1395. https://doi.org/10.1016/10.1007/s10902-018-9999-6.

Hills, P., & Argyle, M. (2002). The Oxford happiness questionnaire: A compact scale for the measurement of psychological well-being. *Personality and Individual Differences, 33*(7), 1073–1082. https://doi.org/10.1016/S0191-8869(01)00213-6.

Höfer, S., Hausler, M., Huber, A., Strecker, C., Renn, D., & Höge, T. (2019). Psychometric characteristics of the German values in action inventory of strengths 120-item short form. *Applied Research in Quality of Life*, 1–15. doi: https://doi.org/10.1007/s11482-018-9696-y, 15, 597, 611.

Hubley, A. M., & Zumbo, B. D. (2011). Validity and the consequences of test interpretation and use. *Social Indicators Research, 103*(2), 219–230. https://doi.org/10.1007/s11205-011-9843-4.

Huta, V., & Ryan, R. M. (2010). Pursuing pleasure or virtue: The differential and overlapping well-being benefits of hedonic and Eudaimonic motives. *Journal of Happiness Studies, 11*(6), 735–762. https://doi.org/10.1007/s10902-009-9171-4.

Iannello, P., Sorgente, A., Lanz, M., & Antonietti, A. (2021). Financial well-being and its relationship with subjective and psychological well-being among emerging adults: Testing the moderating effect of individual differences. *Journal of Happiness Studies, 22*, 1385–1411. https://doi.org/10.1007/s10902-020-00277-x.

Joseph, S., & Wood, A. (2010). Assessment of positive functioning in clinical psychology: Theoretical and practical issues. *Clinical Psychology Review, 30*(7), 830–838. https://doi.org/10.1016/j.cpr.2010.01.002.

Judge, D. (2018). *Evaluating a Trauma-Informed Resilience Curriculum in a Public Alternative High School: Student Treatment Outcomes and Staff Perceptions of Implementation* (Doctoral dissertation). Retrieved from https://digital.lib.washington.edu/researchworks/handle/1773/42103

Kállay, É., & Rus, C. (2014). Psychometric properties of the 44-item version of Ryff's psychological well-being scale. *European Journal of Psychological Assessment, 30*(1), 15–21. https://doi.org/10.1027/1015-5759/a000163.

Kern, M. L., Waters, L. E., Adler, A., & White, M. A. (2015). A multidimensional approach to measuring well-being in students: Application of the PERMA framework. *Journal of Positive Psychology, 10*(3), 262–271. https://doi.org/10.1080/17439760.2014.936962.

Keyes, C. L. M., Shmotkin, D., & Ryff, C. D. (2002). Optimizing well-being: The empirical encounter of two traditions. *Psycnet.Apa.Org, 82*(6), 1007–1022. https://doi.org/10.1037/0022-3514.82.6.1007.

Little, T. (2013). *The Oxford handbook of quantitative methods: Foundations* (Vol. 1). Oxford University press.

Liu, Y., Li, Z., & Liu, H. (2019). Reporting valid and reliable overall scores and domain scores using bi-factor model. *Article Applied Psychological Measurement, 43*(7), 562–576. https://doi.org/10.1177/0146621618813093.

Margola, D., Fenaroli, V., Sorgente, A., Lanz, M., & Costa, G. (2019). The family relationships index (FRI): Multilevel confirmatory factor analysis in an Italian community sample. *European Journal of Psychological Assessment, 35*(3), 335–345. https://doi.org/10.1027/1015-5759/a000427.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw Hill.

Peralta-Pizza, F., Pinzón, D. C., Gaitán, H. G., Eslava-Schmalbach, J., & Rodriguez-Malagon, N. (2019). Google scholar to identify research studies. *Cochrane Database of Systematic Reviews*. https://doi.org/10.1002/14651858.mr000049.

Rahm, T., & Heise, E. (2019). Teaching happiness to teachers-development and evaluation of a training in subjective well-being. *Frontiers in Psychology, 10*, 2703. https://doi.org/10.3389/fpsyg.2019.02703.

Reebs, A., Yuval, K., Hadash, Y., Gebremariam, S. G., & Bernstein, A. (2020). *Mindfulness-Based Trauma Recovery for Refugees (MBTR-R): Randomized Waitlist-Control Evidence of Efficacy and Safety.* Retrieved from https://mindrxiv.org/jyhwa/download?format=pdf

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

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

Sayans-Jiménez, P., Cuadrado, I., Rojas, A. J., & Barrada, J. R. (2017). Extracting the evaluations of stereotypes: Bi-factor model of the stereotype content structure. *Frontiers in Psychology, 8*, 1692. https://doi.org/10.3389/fpsyg.2017.01692.

Schrank, B., Riches, S., Coggins, T., Tylee, A., & Slade, M. (2013). From objectivity to subjectivity: Conceptualization and measurement of well-being in mental health. *Neuropsychiatry, 3*(5), 525–534. https://doi.org/10.2217/NPY.13.58.

Seligman, M. E. P., & Csikszentmihalyi, M. (2014). Positive psychology: An introduction. In *Flow and the foundations of positive psychology* (pp. 279–298). Springer, Dordrecht. doi: https://doi.org/10.1007/978-94-017-9088-8_18.

Seligman, M. (2018). PERMA and the building blocks of well-being. *Journal of Positive Psychology, 13*(4), 333–335. https://doi.org/10.1080/17439760.2018.1437466.

Siddaway, A. P., Wood, A. M., & Hedges, L. V. (2019). How to do a systematic review: A best practice guide for conducting and reporting narrative reviews, meta-analyses, and meta-syntheses. *Annual Review of Psychology, 70*, 747–770. https://doi.org/10.1146/annurev-psych-010418-102803.

Singh, K., & Bandyopadhyay, S. (2020). Enhancing college students well-being: The psycho-spiritual well-being intervention. *Journal of Human Behavior in the Social Environment*, 1–22. https://doi.org/10.1080/10911359.2020.1823294.

Sirigatti, S., Stefanile, C., Giannetti, E., Iani, L., Penzo, I., & Mazzeschi, A. (2009). Assessment of factor structure of Ryff's psychological well-being scales in Italian adolescents. *Bollettino di Psicologia Applicata, 259*(56), 30–50.

Sorgente, A., & Lanz, M. (2019). The multidimensional subjective financial well-being scale for emerging adults: Development and

validation studies. *International Journal of Behavioral Development, 43*(5), 466–478. https://doi.org/10.1177/0165025419851859.

Sorgente, A., Tagliabue, S., Andrade, C., Oliveira, J. E., Duan, W., & Lanz, M. (2020). Gender, age, and cross-cultural invariance of brief inventory of thriving among emerging adults. *Measurement and Evaluation in Counseling and Development. online first*, 1–16. https://doi.org/10.1080/07481756.2020.1827434.

Sorgente, A., Tagliabue, S., & Lanz, M. (2019). Brief inventory of thriving: Testing Italian version's reliability using SEM. *Psicologia Della Salute, 2019*(1), 97–117. https://doi.org/10.3280/PDS2019-001006.

Springer, K. W., & Hauser, R. M. (2006). An assessment of the construct validity of Ryff's scales of psychological well-being: Method, mode, and measurement effects. *Social Science Research, 35*(4), 1080–1102. https://doi.org/10.1016/j.ssresearch.2005.07.004.

Stiglitz, J. E., Sen, A., & Fitoussi, J. P. (2009). *The measurement of economic performance and social progress revisited: Reflections and Overview.* Retrieved from https://hal-sciencespo.archives-ouvertes.fr/file/index/docid/1069384/filename/wp2009-33.pdf

Su, R., Tay, L., & Diener, E. (2014). The development and validation of the comprehensive inventory of thriving (CIT) and the brief inventory of thriving (BIT). *Applied Psychology: Health and Well-Being, 6*(3), 251–279. https://doi.org/10.1111/aphw.12027.

Sultanoff, S. M. (1997). Treating the whole person: A holistic approach to psychotherapy. *The Humanistic Psychologist, 25*(2), 126–137. https://doi.org/10.1080/08873267.1997.9986876.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Pearson.

Ugwu, C., Iorfa, S. K., Ogba, K. T., Effiong, J. E., & Ugwu, T. O. (2018). Thriving among married rural Nsukka women of Igbo extraction: Contributions of domestic violence and psychological distress. *Nigerian Journal of Psychological Research, 14*, 25–33. https://doi.org/10.4103/1119-3077.113449.

Voorhees, C. M., Brady, M. K., Calantone, R., & Ramirez, E. (2016). Discriminant validity testing in marketing: An analysis, causes for concern, and proposed remedies. *Journal of the Academy of Marketing Science, 44*(1), 119–134. https://doi.org/10.1007/s11747-015-0455-4.

Watkins, M. S., Teh, L. A., & Fernandez, K. T. G. (2019). Exploring the use of therapeutic horticulture to enhance the psychological well-being of female survivors of childhood sexual abuse. *Journal of Therapeutic Horticulture, 29*(1), 25–38.

Wiese, C. W., Tay, L., Su, R., & Diener, E. (2018). Measuring thriving across nations: Examining the measurement equivalence of the comprehensive inventory of thriving (CIT) and the brief inventory of thriving (BIT). *Applied Psychology: Health and Well-Being, 10*(1), 127–148. https://doi.org/10.1111/aphw.12119.

Willis, G. B. (2004). *Cognitive interviewing: A tool for improving questionnaire design.* New York, NY: Sage.

Zaiţ, A., & Bertea, P. S. P. E. (2011). Methods for testing discriminant validity. *Management & Marketing Journal, 9*(2), 217–224.

Zumbo, B. D. (2005). Structural Equation Modeling and Test Validation. In Brian Everitt and David C. Howell, *Encyclopedia of Statistics in Behavioral Science,* (pp. 1951-1958). Chichester, UK: John Wiley & Sons Ltd. doi: https://doi.org/10.1002/0470013192.bsa654.