RESEARCH ARTICLE

Addressing the construct validity of the "individual problems and strengths" scale [version 1; peer review: 1 approved with reservations]

Rune Zahl-Olsen[1], Nicolay Gausel[2], Åshild Tellefsen Håland[1], Terje Tilden[3]

[1]Department of Child and Adolescent Mental Health, Sørlandet sykehus, Kristiansand, Agder, 4615, Norway
[2]Faculty of Social Sciences, University of Stavanger, Stavanger, Rogaland, Norway
[3]Modum Bad Research Institute, Modum Bad, Vikersund, Norway

v1 **First published:** 03 Oct 2022, **11**:1129
https://doi.org/10.12688/f1000research.125176.1
Latest published: 03 Oct 2022, **11**:1129
https://doi.org/10.12688/f1000research.125176.1

Abstract

Background: Routine Outcome Monitoring (ROM) systems have been used to monitor how a client's life changes over the course of therapy. However, if a ROM system is to be used, the system should have sufficient construct validity to warrant its usage. In the current study we sought to test the construct of the "individual problems and strengths" (IPS) measurement scale, a sub-section of the "Systemic Therapy Inventory of Change" (STIC).
Methods: We used a factorial construct validation procedure utilizing a stepwise confirmatory factor analysis approach on a sample of 841 clients of couple and family therapy.
Results: We found support for the original "8-factor" version of the IPS but failed to find support for the "1-factor" version and the "higher order factor structure".
Conclusions: The investigation uncovered that the measurement tool is still under development and since the factorial construct (and the scale-reliability) was only supported for the original "8-factor" model, we encourage a pause in administering the IPS in clinical practice.

Keywords

Validity, Reliability, Confirmatory Factor Analysis, Cronbach, Routine Outcome Monitoring, STIC, IPS, Measurement, Therapy, Feedback instruments, Clinical tool

Open Peer Review

Approval Status ?

	1
version 1 03 Oct 2022	? view

1. **Peter Stratton**, University of Leeds, Leeds, UK

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Rune Zahl-Olsen (Rune.Zahl-Olsen@sshf.no)

Author roles: Zahl-Olsen R: Conceptualization, Data Curation, Formal Analysis, Investigation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Gausel N**: Conceptualization, Formal Analysis, Methodology, Supervision, Validation, Writing – Review & Editing; **Tellefsen Håland Å**: Project Administration, Supervision, Writing – Review & Editing; **Tilden T**: Conceptualization, Data Curation, Project Administration, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This project was funded by Sorlandet Kompetansefond and Sparebanken Sor, Norway (Grant no: 2017/24).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Zahl-Olsen R, Gausel N, Tellefsen Håland Å and Tilden T. **Addressing the construct validity of the "individual problems and strengths" scale [version 1; peer review: 1 approved with reservations]** F1000Research 2022, **11**:1129 https://doi.org/10.12688/f1000research.125176.1

First published: 03 Oct 2022, **11**:1129 https://doi.org/10.12688/f1000research.125176.1

Introduction

Even though psychotherapy has been found to be effective for a large portion of clients seeking help (Sexton *et al.*, 2013), not all benefit from therapy; in fact, some even get worse (Ogles, 2013). Attempting to better understand and explain why some clients benefit while others do not, quantitative systems consisting of standardized questionnaires (i.e., Routine Outcome Monitoring systems – "ROM") have been developed over the last years (Lambert, 2010; Ogles, 2013; Tilden & Wampold, 2017). These ROM systems tap into various aspects of the client's life, and if regularly responded to, the calculated mean score of these aspects will enable the therapist (and the client) to monitor change over the course of therapy (Duncan *et al.*, 2004; Zahl-Olsen & Oanes, 2017). However, to administer a ROM system to monitor change, one needs reassurance that the ROM system is measuring what it is supposed to be measuring (i.e., its construct validity) and that the results of the measurement can be trusted (i.e., its scale-reliability). One of the ROM systems increasingly applied within the field of psychotherapy, and especially within couple and family therapy settings (Tilden & Wampold, 2017), is the *"Systemic Therapy Inventory of Change"* (STIC) (Pinsof *et al.*, 2009).

The STIC addresses therapeutic change through a battery of six larger thematic sets of questionnaires, where one of these, the clinically important *"individual problems and strengths"* (IPS), taps into 8 typical aspects of distress and everyday difficulties that clients experience as they go through therapy. These include, for example, the degree to which the client can change plans and cope with changes (i.e., *Flexibility/resilience*), being able to express and share feelings (i.e., *Open expression*), managing daily tasks such as work and household (i.e., *Life functioning*), or coping with negative emotional states (i.e., *Negative affect*).

The 8-factorial structure of the IPS was first validated by the developers of STIC (Pinsof *et al.*, 2009) on 188 clients seeking outpatient therapy in the Chicago area using confirmatory factor analysis (CFA). The factorial model yielded a good fit of data (The Root Mean Square Error of Approximation (RMSEA) = .06 and Comparative Fit Index (CFI) = .94) with low to good reliability for the 8 factors (α ranging from .54 to .89). Some years later, the same team (Pinsof *et al.*, 2015) tried to replicate their "8-factorial" IPS structure, only now with more statistical power increasing their sample to 581 clients. However, this time the CFA did not fare so well with the "8-factorial" structure. It received poor fit of data (RMSEA = .117 and CFI = .59) with low to good reliability (α ranging from.45 to.85). In this paper, they also designed a factorial structure where all items loaded onto a global "single factor" (i.e., a "1-factor" solution) of IPS, arguing that all items were essentially a representation of the same psychological experience of individual problems and strengths. However, like the 8-factorial structure, this "1-factorial" structure also fit the data poorly (RMSEA = .099 and CFI = .71). Consequently, Pinsof and colleagues proposed a factorial solution they termed a "higher order factor structure." This was a CFA where all items were loaded onto the 8 individual factors and then loaded onto a global "higher-order" IPS factor. This factorial solution yielded a borderline decent fit of data (RMSEA = .069 and CFI = .86) which Pinsof *et al.* (2015) viewed as a validation of their factorial solution.

In 2018, the same research team, only now led by Zinbarg *et al.* (2018), addressed the IPS and its relationship with other ROMs based on reports from 476 outpatient clients in the Chicago area. This time, however, they reported means and standard deviations for the 8 different factors along with their reliabilities (α ranging from .66 to .87) but failed to report any analysis or statistical support for an 8-factorial structure merely stating that the IPS "load on eight factors" (p.737). In our view, it is unclear why they did not report the statistical results of the original "8-factorial" solution but reported results from CFA on a "2-factor" solution in which they received mixed to fairly good support (RMSEA = .111 and CFI = .98) and on a "1-factor" solution which provided almost identical results (RMSEA = .115 and CFI = .98).

Taken together, despite its growing popularity, the factorial structure of STIC has been replicated only twice, and its clinically important "8-factor" solution of IPS has been validated only once, in the original 2009 paper (Pinsof *et al.*, 2009). The failure to replicate the 8-factorial structure in 2015 (Pinsof *et al.*, 2015) and the lack of reporting CFA results on the "8-factor" structure in 2018 (Zinbarg *et al.*, 2018), along with mixed levels of scale-reliability, all indicate that construct validation (and development) of the IPS is an ongoing and continuous process. Moreover, to date, no attempts to replicate the IPS have been published on samples outside the US or in a language other than English.

The current study

In the current study, we aimed to see if we could replicate the factorial structure of the IPS (Pinsof *et al.*, 2009) in a Norwegian cultural context using the Norwegian language. Specifically, it would mean we would first test the hypothesis that IPS represents an "8-factor" solution. Secondly, as Pinsof *et al.* (2009) suggest, the IPS can be collapsed into a "1-factor" solution, so we decided to test this "1-factor" structure. Finally, as Pinsof and colleagues presented a "higher order factor structure" in 2015, we decided to test this alternative factor solution as well.

To test these factorial structures, we decided to follow the stepwise approach to construct validity as laid out by Gausel and colleagues (Gausel *et al.*, 2012, 2016, 2018; Pardede *et al.*, 2021). This approach recommends a four-step procedure

where the preferred hypothesized factorial structure serves as a point of reference (in our case, the original 8-factorial structure). All other factorial solutions are compared against the preferred hypothesized model and each other. The best-fitting model (which should be the hypothesized model) would "win". However, should the hypothesized model "lose" (i.e., fail to achieve the best fit), then more exploration of data is needed; ideally using exploratory factor analyses for the latter to return to a revised, hypothesized model to be tested with the same Gausel *et al.*, procedure (for a discussion and practical example of this approach, see Pardede *et al.* (2021)).

Our aim to test and attempt to replicate these various factors is primarily motivated by the fact that the IPS is increasingly deployed in various clinical settings. For instance, the 8-factorial solution is used by therapists to calculate eight different individual mean values used to monitor change throughout therapy, and the 1-factor solution is used to calculate an overall mean value to trace change more easily throughout therapy. Naturally, if we fail to replicate the factorial structure of the increasingly deployed IPS in therapy, therapists should be informed that the ROM they are using to interpret and trace change in client's life throughout the therapy is flawed. Thus, it is of great importance for therapists to know this, but most of all for the well-being of our clients.

Methods

This study's data were collected in Norway, where couple and family therapy treatment is offered to the general public in stepped levels of care. Two agencies provide the initial level of care, for which no referral is required. An outpatient agency represents the second level of care, for which a referral is required. A referral is required for the third and final level of care, which is represented by an inpatient facility. Since the data is derived from standard clinical practice, no inclusion or exclusion criteria were applied, other than the criteria each site uses to accept patients for treatment. The data was collected from clients of over 40 therapists at all three levels of couple and family therapy in Norway through online questionnaires[1].

Ethics and consent

The PhD work that led to this manuscript was approved by the Modum Bad Ombudsman for Data Protection and the Regional Ethics Committee for Medical Research (2017/96/REK Sør-øst C, approved March 6, 2017). The primary study was also approved by the Modum Bad Ombudsman for Data Protection and the Regional Ethics Committee for Medical Research with human subjects; the pilot study was approved Nov. 13, 2009 (2009/927/REK Sør-øst C) and the RCT study was approved Jun. 13, 2016. Written informed consent was obtained from participants. For participants under the age of 16, consent was obtained from parents or guardians. This study investigated data from one of the questionnaires used in both a multi-site RCT study investigating the effects of the use of online feedback in therapy, registered at ClinicalTrials.gov (NCT01873742), as well as from a prior pilot study (Tilden *et al.*, 2015). Ethical recommendations have been followed.

Participants and procedure

The RCT and pilot studies recruited 841 clients (51.8% women; mean age: 40; age range: 12-72) through ordinary clinical practice from March 2010 to April 2016. Written informed consent was obtained from each participant. Data is available for download (Zahl-Olsen *et al.*, 2021a) and the variables and the data are described in more detail by Zahl-Olsen *et al.* (2021b). Data was collected as the clients began their therapeutic process.

Measures

The original "individual problems and strengths" (IPS) thematic subscale (Pinsof *et al.*, 2009) has a total of 22 items theorized to tap into 8 different subscales (see Table 1 for correlations and descriptive statistics). *Flexibility/resilience* consisted of three items (α = .67) measured with a scale ranging from 1 (1st item: "very easy", 2nd and 3rd items "strongly disagree") to 5 (1st item: "very hard", 2nd and 3rd items "completely disagree"): "How easy is it for you generally to overcome difficulties?", "When what I'm trying doesn't work out, I can change my approach or my plans" and "When I get upset, I find healthy ways to make myself feel better". *Life functioning* measure consisted of two items (α = .78) measured with a scale ranging from 1 ("really bad") to 5 ("really good"): "Performing work/school/household tasks" and "Managing day-to-day life". *Open expression* consisted of two items (α = .68) measured with a scale ranging from 1 ("completely disagree") to 5 ("completely agree"): "I can openly express my feelings" and "I can speak up for myself when the situation calls for it". *Self-acceptance* consisted of two items (α = .71) measured with a scale ranging from 1 ("completely disagree") to 5 ("completely agree"): "I can be myself in every situation" and "I am comfortable with who I am". *Disinhibition* consisted of three items (α = .53) measured with a scale ranging from 1 ("never") to 5 ("all the time"):

[1]The current manuscript originates from the first author's PhD-thesis. The thesis consisted of a theoretical introduction with three different articles presenting findings of the thesis. The current manuscript is one of these three, only reworked and modified from how it was originally presented in the thesis.

Table 1. Scale inter-correlations and descriptive statistics.

	Variable	1	2	3	4	5	6	7	8
1	Flexibility/resilience	-							
2	Life functioning	.42*	-						
3	Open expression	.41*	.23*	-					
4	Self-acceptance	.61*	.40*	.49*	-				
5	Disinhibition	.37*	.35*	.13*	.36*	-			
6	Negative affect	.50*	.56*	.19*	.52*	.47*	-		
7	Self-misunderstanding	.47*	.36*	.31*	.48*	.36*	.43*	-	
8	Substance abuse	.15*	.16*	.03	.13*	.29*	.21*	.14*	-
	Mean	3.48	3.35	3.65	3.14	4.51	3.43	3.51	4.81
	SD	.76	.83	1.01	1.10	.53	.79	1.08	.35

*p<.001.

"Thought about seriously harming or killing someone", "Had fits of rage you could not control", and "Had urges or impulses that you could not control". *Negative affect* consisted of six items (α = .86) measured with a scale ranging from 1 ("never") to 5 ("all the time"): "Had thoughts or images over and over again that you could not get rid of", "Felt tense or anxious", "Felt sad most of the day", "Thought about ending your life", "Felt hopeless about the future", and "Not enjoyed things as much as you used to". *Self-misunderstanding* consisted of two items (α = .66) measured with a scale ranging from 1 ("completely disagree") to 5 ("completely agree"): "I don't understand why I do the things I do" and "It's tough for me to know what I'm feeling". *Substance abuse* consisted of two items (α = .20) measured with a scale ranging from 1 ("never") to 5 ("all the time"): "Drank too much alcohol" and "Used illegal drugs/misused prescribed medication". The original English STIC version was translated into Norwegian according to the procedures outlined by Wild *et al.* (2005), which included preparation, forward translation by two independent interpreters, reconciliation, back translation, back translation review, harmonization, cognitive debriefing, and finalization. No test of the reliability of the back translation was performed before implementation in this study.

Statistical analysis

We used *AMOS 25* from *IBM* to test our hypothesis with a confirmatory factor analysis (CFA) using maximum likelihood estimation. We adopted Gausel *et al.*' (2012, 2016, 2018) stepwise "construct validity" (p. 945) approach by first testing the fit of the preferred model. Then, in a second step, testing the fit of the competing model. In a third step, comparing the fit of the competing model up against the preferred model, and finally, in a fourth step, by comparing the fit of the preferred model up against the fit of other plausible alternatives. In line with the recommendations of Gausel *et al.* (2012, 2016), we first tested the preferred "8-factor" model. Then, in the second step, we tested the "1-factor" model. In the third step, we compared the fit of the "8-factor" model against the fit of the "1-factor" model. In the fourth step, we compared the "8-factor" model against other plausible, data-driven alternatives, as well as the "higher order factor structure" as suggested by Pinsof *et al.* (2015). In line with recommendations by Gausel and colleagues, latent factors were allowed to correlate, but no items were allowed to cross-load on any of the factors, and no error terms were allowed to correlate.

Results

Step 1: Attempting to validate the "8-factor" construct of IPS

In a first step, we tested the "8-factor" version of the IPS. Despite a significant chi-square, χ^2 (181) = 605,994, p <.001, χ^2/df = 3.35, the "8-factor" model fit the data well as indicated by the other fit-indices: *IFI* = .936, *CFI* = .935, *RMSEA* = .053 [.048 - .058], *AIC* = 793.994. As seen in Figure 1, all factor loadings were significant (all p values < .001), ranging from standardized λ = .30 to .86, with most above .55 or higher indicating that factors were well defined by their respective items (Gausel *et al.*, 2012, 2016, 2018). The correlations among the eight different factors ranged from low (r = .06, p = .398) to high (r = .89, p < .001) with more than half of the items producing an explained variance of around 50% reaching an ideal level for a CFA (Kline, 2016).

Step 2: Attempting to validate the "1-factor" construct of IPS

In a second step, we tested the "1-factor" version of the IPS. Here, all items are theorized to be representative of the single construct. As such, we allowed all items to load onto a single "IPS factor". This model fit the data poorly as indicated by a

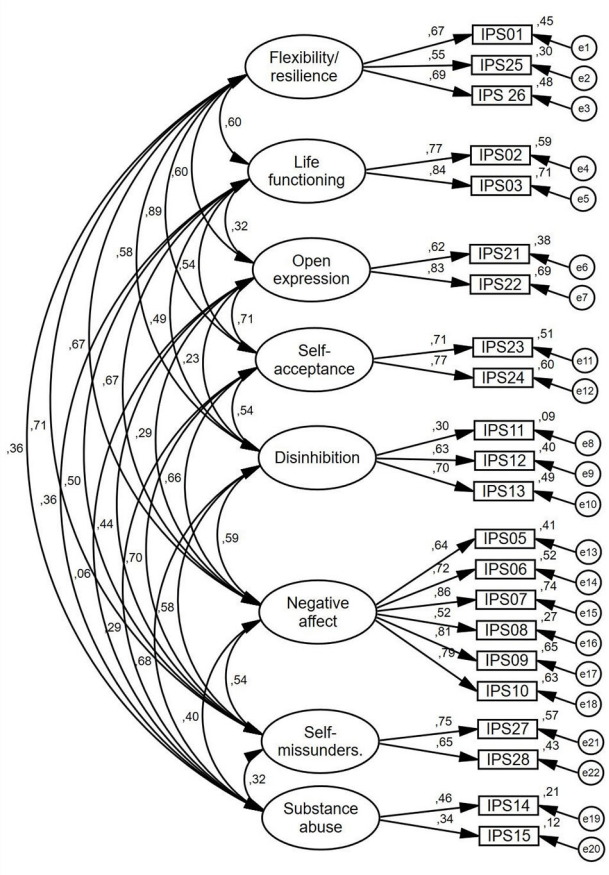

Figure 1. The "8-factor" model.

significant chi-square, χ^2 (209) = 1949,931, p <.001, a very high χ^2/df = 9.33, very low *IFI* = .736, very low *CFI* = .734, and a high *RMSEA* = .100 [.096 - .104], *AIC* = 2081.93. Not only did this "1-factor" model represent the data poorly, only half of the 22 items had factor loadings larger than λ .55, and only three of the items in the undifferentiated model reached the suggested 50% level of explained variance (Kline, 2016).

Step 3: Comparing the "8-factor" version against the "1-factor" version

Even though the fit indices clearly communicated that the "1-factor" model should be rejected, we decided to continue to follow Gausel *et al.*' (2012, 2016, 2018) recommendations comparing the "1-factor" model up against the "8-factor" model. The "8-factor" model fit data significantly better than the "1-factor" model; Δ χ^2 (28) = 1343.937, p < .001. Moreover, the difference in AIC was substantial, Δ *AIC* = 1287.937, demonstrating that the "8-factor" model was indeed superior to the "1-factor" model.

Step 4: Comparing the "8-factor" version of IPS up against other plausible data-driven alternatives, as well as the "higher order factor structure"

In the final step of the recommendations by Gausel *et al.* (2012, 2016, 2018), we compared the "8-factor" version of IPS up against other meaningful alternatives. Looking at the four correlations in Figure 1 that are higher than or equal to r = .70, we identified five alternative models that represented plausible data-driven alternatives to the 8-factor model. However, the "8-factor" model proved superior to all these alternative models. First, it fit better than a model collapsing "open expression" with "self-acceptance", Δ χ^2 (7) = 149.44, p < .001 and Δ *AIC* = 135.440. Second, it fit better than a model collapsing "flexibility/resilience" with "self-acceptance", Δ χ^2 (7) = 27.761, p < .001 and Δ *AIC* = 13.761. Third, it fit better than a model collapsing "flexibility/resilience" with "self-misunderstanding," Δ χ^2 (7) = 75.735, p < .001 and Δ *AIC* = 61.735. Forth, it fit better than a model collapsing "self-misunderstanding" with "self-acceptance." Δ χ^2 (7) = 95.713, p < .001 and Δ *AIC* = 81.713. Finally, it fit better than a model where all four mentioned factors were collapsed, Δ χ^2 (18) = 301.710, p < .001, and Δ *AIC* = 260.71.

In 2015, Pinsof and colleagues tested a so-called "higher order factor structure" where they allowed their individual factors *"like IPS Negative Affect and Open Expression"* to load onto *"a single second-order general factor like IPS"* (p.470). They underlined that this analysis was *"theoretically and methodologically important because finding support for a higher order model (…) demonstrates that each of the scale's group factors links to a single higher order factor that underlies the scale and its factors"* (p. 470). Consequently, we decided to test the "higher order factor structure" where each of the 8-factors is allowed to load independently onto a "higher-order" IPS-factor. This model approached a borderline acceptable fit of data, χ^2 (201) = 818,833, $p < .001$, $\chi^2/df = 4.07$, *IFI* = .906, *CFI* = .906, *RMSEA* = .060 [.056 - .065], *AIC* = 966.833. However, when we compared this "higher order" model up against the original "8-factor" model, the "8-factor" model fit data significantly better than the "higher order" model; $\Delta\chi^2$ (20) = 212.839, $p < .001$, with a large difference in AIC; Δ *AIC* = 212.839. Hence, the original "8-factor" model was superior to the suggested "higher order factor structure".

Discussion

Over recent years, research on psychotherapy has focused on how best to monitor clients' change throughout therapy using ROMs — Routine Outcome Monitoring (Duncan *et al.*, 2004; Lambert, 2010; Pinsof *et al.*, 2009). Naturally, it is important for a therapist (and a researcher) to be reassured that the ROM administered will represent the clients' experiences in the best way possible. One of the ROMs believed to do this just so is the STIC system developed by Pinsof *et al.* (2009). However, the STIC system has been validated only twice, and the highly clinically-relevant sub-section, the IPS, has been successfully validated only once, only in the Chicago area and only in the English language. Thus, there were ample grounds to test the construct validity of the IPS and do so in a different culture and a different language.

In the current study, we employed Gausel *et al.*' (2012, 2016, 2018) stepwise approach to "construct validation." The advantage of their approach is that it is a clear-cut, step-by-step construct validation procedure where one can test a factorial structure up against other factorial structures in order to establish which is the best fitting model. Our first step was to test the fit of the original "8-factor" model as suggested by Pinsof *et al.* (2009). This first step was important because if it failed to fit well, it would be pointless to compare it to any other models (2018; 2012; 2016). As expected, the "8-factor" model *did* represent the data well. In fact, whichever way we tried to modify the combination of factors in the different steps of the CFA approach, the "8-factor" model always came out as the superior factorial solution. By such, the step-by-step analysis provided validating support to the original "8-factor" model as developed by Pinsof *et al.* (2009), and it supports Zinbarg *et al.*' (2018) argumentation that a multi-faceted version of the IPS (and STIC) would provide the most accurate information about the complex lives of clients.

We also tested the proposed "1-factor" model (Pinsof *et al.*, 2009) and the "higher order factor structure" (Pinsof *et al.*, 2015). In terms of the "1-factor" model, we found it to be representing the data very poorly. In fact, when we compared this model up against the "8-factor" model, the "1-factor" proved to be inferior in all ways. This result goes against Pinsof *et al.*' (2009, 2015) argumentation that all items in the IPS are indifferently representative of the overall construct, and it goes against the common therapeutic practice to calculate an overall mean in order to trace change in therapy (Oanes *et al.*, 2015; Spanier, 1988). In terms of the "higher order factor structure," we found it to achieve borderline acceptable fit. However, as it fit significantly worse than our "8-factor" model, the "higher order factor structure" was found to be an inferior alternative to the better fitting "8-factor" model.

In terms of scale-reliability, only one of the 8 IPS sub-scales achieved a Cronbach's alpha more than .80 (i.e., Negative Affect), two achieved more than .70 (i.e., Life functioning and Self-acceptance), three more than .60 (i.e., Flexibility/ resilience, Open expression, and Self-misunderstanding), one more than.50 (i.e., Disinhibition), and one barely made it to.20 (i.e., Substance abuse). Clearly, a measurement model obtaining low to mainly moderate and acceptable levels of scale-reliability levels indicates a need for further development of the various measurement scales (Schmitt, 1996). That said, finding levels of scale-reliability to be low in our study did not come as a surprise. As earlier communicated by the developers of the measurement tool (Pinsof *et al.*, 2009), the low reliability of their scales constitutes *"a major methodological concern"* (p. 151). Due to this, they aimed *"to increase the reliability of the subscales with low alphas"* (p. 151) in future studies. However, six years later, Pinsof *et al.* (2015) still struggled with low reliabilities, reiterating that it *"is concerning"* (p. 478) that so many subscales of their measurement tool suffer from poor levels of scale-reliability. In light of our study, we cannot but agree: it *is* concerning, especially as the IPS (and STIC) is increasingly used in clinical practice to measure and monitor how clients' change over the course of therapy (Pinsof *et al.*, 2015).

Possible limitations

We have to admit that the current study tested only one (the IPS) of six sub-themes within a comprehensive ROM system, the STIC (Pinsof *et al.*, 2009). Therefore, we cannot say much in terms of the remaining other sub-themes but encourage future testing of the remaining five. Moreover, we are unable to have opinions about the construct validity or scale

reliability of other ROM systems, such as the Systemic Clinical Outcome and Routine Evaluation (Carr & Stratton, 2017). Nevertheless, we generally encourage developers and independent researchers to test other ROM systems' construct validity and scale reliability. After all, the focus should be on the client and how best to understand and care for her/him as they go through therapy. This demands a measurement tool that can be trusted.

Conclusion

Taken together, our study supports an "8-factor" solution of the IPS as originally developed by Pinsof *et al.* (2009) both cross-culturally (in the Norwegian culture) and cross-linguistically (the Norwegian language). We see this as a major step forward in construct validation of their IPS outside the US and the English language. However, our positivity comes with a fair amount of soberness. As we see it, the IPS may have achieved support for its "8-factorial" structure, but it has failed to live up to acceptable standards in terms of scale reliability. We have sympathy for the developer's eagerness to put the measurement tool into practice, but as long as the tool *is under development,* we call on developers to pause the practical use of it until more research has clarified its construct validity and, importantly, increased its scale reliability.

Data availability

Underlying data

Mendeley Data: Dataset of the Individual Problems and Strengths scale (IPS): A clinical sample, https://doi.org/10.17632/fk5v8n726c.1 (Zahl-Olsen, Tellefsen Haaland and Tilden, 2021a).

This project contains the following underlying data:

- Data_IPS.sav
- Data_IPS.csv

Extended data

The data key for these files can be found in: Zahl-Olsen, R., Haaland, A. T., & Tilden, T. (2021b). Data on the individual problems and strengths scale from the systemic therapy inventory of change. Clinical samples from Norway. *Data in Brief*, *39*. https://doi.org/10.1016/j.dib.2021.107577

References

Carr A, Stratton P: **The SCORE family assessment questionnaire: a decade of progress.** *Fam. Process.* 2017; **56**(2): 285–301.
PubMed Abstract | Publisher Full Text

Duncan BL, Miller SD, Sparks J: *The heroic client: a revolutionary way to improve effectiveness through client-directed, outcome-informed therapy.* Rev. ed. Jossey-Bass; 2004.

Gausel N, Leach CW, Mazziotta A, *et al.*: **Seeking revenge or seeking reconciliation? How concern for social-image and felt shame helps explain responses in reciprocal intergroup conflict.** *Eur. J. Soc. Psychol.* 2018; **48**(1): O62–O72.
Publisher Full Text

Gausel N, Leach CW, Vignoles VL, *et al.*: **Defend or repair? Explaining responses to in-group moral failure by disentangling feelings of shame, rejection, and inferiority.** *J. Pers. Soc. Psychol.* 2012; **102**(5): 941–960.
Publisher Full Text

Gausel N, Vignoles VL, Leach CW: **Resolving the paradox of shame: Differentiating among specific appraisal-feeling combinations explains pro-social and self-defensive motivation.** *Motiv. Emot.* 2016; **40**(1): 118–139.
Publisher Full Text

Kline RB: *Principles and practice of structural equation modeling.* 4th ed. Guilford publications; 2016.

Lambert MJ: **Yes, it is time for clinicians to routinely monitor treatment outcome.** 2010.
Publisher Full Text

Oanes CJ, Borg M, Karlsson B: **Significant Conversations or Reduced Relational Capacity? Exploring Couple and Family Therapists' Expectations for Including a Client Feedback Procedure.** *Aust. N. Z. J. Fam. Ther.* 2015; **36**(3): 342–355.
Publisher Full Text

Ogles BM: **Measuring change in psychotherapy research.** Lambert MJ, editor. *Bergin and Garfield's Handbook of psychotherapy research and behavior change.* 6th ed. Wiley; 2013; pp. 134–166.

Pardede S, Gausel N, Høie MM: **Revisiting the "the breakfast club": Testing different theoretical models of belongingness and acceptance (and social self-representation).** *Front. Psychol.* 2021; **11**: 3801.
PubMed Abstract | Publisher Full Text

Pinsof WM, Zinbarg RE, Lebow J, *et al.*: **Laying the foundation for progress research in family, couple, and individual therapy: The development and psychometric features of the initial Systemic Therapy Inventory of Change.** *Psychother. Res.* 2009; **19**(2): 143–156.
PubMed Abstract | Publisher Full Text

Pinsof WM, Zinbarg RE, Shimokawa K, *et al.*: **Confirming, Validating, and Norming the Factor Structure of Systemic Therapy Inventory of Change Initial and Intersession.** *Fam. Process.* 2015; **54**(3): 464–484.
PubMed Abstract | Publisher Full Text

Schmitt N: **Uses and abuses of coefficient alpha.** *Psychol. Assess.* 1996; **8**(4): 350–353.
Publisher Full Text

Sexton TL, Datchi C, Evans L, *et al.*: **The effectiveness of couple and family-based clinical interventions.** Lambert MJ, editor. *Bergin and Garfield's handbook of psychotherapy and behavior change.* 6th ed. John Wiley & Sons; 2013; pp. 587–639.

Spanier GB: **Assessing the strengths of the Dyadic Adjustment Scale.** *J. Fam. Psychol.* 1988; **2**(1): 92–94.
Publisher Full Text

Tilden T, Håland ÅT, Hunnes K, *et al.*: **Utprøving av systematisk tilbakemelding i par- og familieterapi: Barrierer og utfordringer.** *Fokus på familien.* 2015; **43**: 292–312.
Publisher Full Text

Tilden T, Wampold BE: *Routine Outcome Monitoring in Couple and Family Therapy: The Empirically Informed Therapist.* Springer; 2017.

Wild D, Grove A, Martin M, *et al.*: **Principles of good practice for the translation and cultural adaptation process for patient-reported outcomes (PRO) measures: report of the ISPOR task force for translation and cultural adaptation.** *Value Health.* 2005; **8**(2): 94–104.
PubMed Abstract | Publisher Full Text

Zahl-Olsen R, Haaland AT, Tilden T: **Dataset of the Individual Problems and Strengths scale (IPS): A clinical sample.** *Mendeley Data.* 2021a; V1.
Publisher Full Text

Zahl-Olsen R, Haaland AT, Tilden T: **Data on the individual problems and strengths scale from the systemic therapy inventory of change. Clinical samples from Norway.** *Data Brief.* 2021b; **39**: 107577.
PubMed Abstract | Publisher Full Text

Zahl-Olsen R, Oanes CJ: **An Anthill of Questions that Made Me Prepare for the First Session: A Clinical Vignette of the Usage of STIC Feedback System.** Tilden T, Wampold BE, editors. *Routine Outcome Monitoring in Couple and Family Therapy.* Springer; 2017; pp. 189–209.
Publisher Full Text

Zinbarg RE, Pinsof WM, Quirk K, *et al.*: **Testing the convergent and discriminant validity of the Systemic Therapy Inventory of Change Initial scales [Empirical paper].** *Psychother. Res.* 2018; **28**: 734–749.
PubMed Abstract | Publisher Full Text

Open Peer Review

Current Peer Review Status: ?

Version 1

Reviewer Report 18 November 2022

https://doi.org/10.5256/f1000research.137454.r152469

?

Peter Stratton
Leeds Institute of Health Sciences (LIHS), University of Leeds, Leeds, UK

The well-written study tackles essential issues in the development and application of ROMs, of reliability and validity. Without good evidence of these, a measure cannot be depended on for general usage. Data from an acceptably large sample were used to investigate the *"individual problems and strengths"* (IPS) subscale of the STIC.

There is no indication of how many in the sample were members of the same couple or family. This would create a risk of non-independence in the sample. Although the items in IPS are individually directed, so not asking about functioning in the relationship, the possibility of correlations within families should be considered.

The statistical procedures are well-referenced and described. They consistently point to the superiority of the 8-factor solution to any measure derived by coalescing the items with the combined measures that will be preferred as a measure of progress in the therapy, showing poor construct reliability. This robust finding suggests that the 8 subscales are measuring different aspects of clients' reports of their lives rather than alternative indicators of any underlying reality. Indicated by the low intercorrelations of Table 1 in which 22 of the 30 coefficients are less than 4.5 so accounting for less than 20% of the variation.

It seems possible that the low Cronbach for substance abuse may have been due to a consistently extreme rating of Mean 4.81 / 5 = 'all the time'. Were nearly all of the sample really drunk or drugged all the time or have I misunderstood?

There is little comment on the potential usefulness to clinicians of a client's scores on the 8 different factors. There is no discussion of how a clinician might make use of the information provided by the IPS. In practice, clinicians often want to make use of the scores of an individual client as an indication of where to focus in the therapy.

The core question addressed by the paper is whether the IPS is useful for a clinician. The conclusion is that, because none of the proposals for creating a summary measure out of the 8

subscales creates an acceptable level of statistical coherence, the IPS is not ready to be used as an outcome measure. This is an important conclusion for an SRM that is being widely promoted.

Typo: End p.6 "*Forth, it fit better than a model...*" - Fourth

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Family therapy; outcome measurement

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.