The Strathclyde Inventory: Development of a brief instrument for assessing outcome in counseling according to Rogers' concept of the fully functioning person

Dr Susan Stephen (corresponding author)
Professor Robert Elliott

Affiliation for all authors at the time the research was conducted: School of Psychological Sciences & Health, University of Strathclyde, Glasgow, United Kingdom

Acknowledgments
We wish to acknowledge the contributions of Elizabeth S. Freire, who initiated the development of the Strathclyde Inventory, Brian Rodgers, who co-developed the SI-16, and Chinatsu Tashiro who carried out an unpublished MSc study based on an early subset of the data used here.

Declaration of interests
No potential competing interest was reported by the authors

Abstract

The Strathclyde Inventory is a self-report instrument assessing Rogers' concept of the fully functioning person. Using data collected from a UK-based counseling service, we investigated its validity for use as an outcome measure, and produced a 12-item brief version that maintained fit to the Rasch model and construct representation.

Keywords: fully functioning person; measure development; outcome measurement; Rasch model; Strathclyde Inventory

The Strathclyde Inventory: Development of a Brief Instrument for Assessing Outcome in Counseling according to Rogers' Concept of the Fully Functioning Person

Counseling researchers continue to be influenced by the work of Carl Rogers. This can be seen in the ongoing development of tests and scales designed for assessment of counseling process, for example, counselor empathic responses (Bayne & Hankey, 2020; Johnson et al., 2021), and change (i.e., psychological growth) predicted by his theory: for example, the Authenticity Scale (Bayliss-Conway et al., 2020), and the Unconditional Positive Self-Regard Scale (Murphy et al., 2020). However, none of these instruments test Rogers' description of the *fully functioning person*: "the [hypothetical] end-point of optimal psychotherapy… the kind of person who would emerge if counseling was maximal" (Rogers, 1963, p. 18).

The Fully Functioning Person

Rogers (1961/1967) identified trends in the process of becoming more fully functioning, initially drawn from his clinical experience. These were: *openness to experience*, the "opposite of defensiveness" (p. 115), allowing the individual to develop greater self-awareness, to integrate a fuller range of feelings and attitudes within their sense of self, and to be more tolerant of ambiguity; *trust in one's organism*, that is, a growing trust in an internal "weighing and balancing" (p. 118) process, enabling the individual to respond to situations in ways that satisfy as many of their intrinsic needs as possible; *an internal locus of evaluation,* increasing acceptance of personal responsibility for making decisions and setting standards for self, rather than seeking approval from others; and *willingness to be a process*, in which the individual moves from an expectation of achieving a fixed outcome (i.e., that their problems have been solved) to the recognition that they are "a fluid process, not a fixed and static entity; a flowing river of change, not a block of solid material; a continually changing constellation of potentialities, not a fixed quantity of traits" (p. 122). In a formal statement of his theory of the fully functioning person, Rogers (1959) proposed two additional characteristics: that they have no conditions of worth, experiencing *unconditional self-regard*; and that they *live in harmony with others* because of the "rewarding character of reciprocal positive regard" (p. 234).

Although Rogers regarded it useful to deconstruct his concept of the fully functioning person for descriptive purposes, he maintained that these characteristics were "quite unitary" (1963, p. 18) in practice, representing a general movement within the person from incongruence to congruence (Rogers, 1959). Indeed, he noted that, as individuals become more fully functioning, they became able to "tolerate a much wider range and variety of feelings, including feelings which were formerly anxiety-producing […by integrating these feelings] into their more flexibly organized personalities" (p. 22) and respond in more constructive, creative ways to difficult life experiences. From this perspective, becoming more fully functioning can be understood as the primary "target", the anticipated outcome of counseling, with the reduction of symptoms of distress an inevitable by-product of the process.

However, as Levitt et al. (2005) identified, self-report instruments regularly used to measure outcome in humanistic counseling tend to be based on medical model goals, such as identifying diagnostic categories and achieving symptom reduction. They described this practice as “akin to weighing oranges with thermometers” (p. 113), and argued for the development and use of instruments that better fit the theoretical concepts of outcome applied in counseling. The Strathclyde Inventory (Freire, 2007) was developed as a response to this perceived gap in humanistic counseling outcome measurement: a self-report instrument designed to measure Rogers’ concept of the fully functioning person.

The Strathclyde Inventory

Freire (2007) extracted descriptions from Rogers’ writings about the fully functioning person (1959, 1961/1967) that captured these six characteristics. Like Rogers, Freire did not conceptualize these characteristics as conceptually distinct; but, rather, interwoven strands within a unitary experience. Freire developed a pilot 51-item instrument with a five-category rating scale, named the Strathclyde Inventory (SI). Thirty of the items were positively worded, and the remaining 21 items were expressed negatively and reverse scored. Higher scores reflected a more fully functioning response. The SI was tested with data collected from a convenience non-clinical sample (trainee and practicing counselors). The results indicated that the SI had excellent score reliability ($\alpha = .94$), and was not substantially associated with social desirability ($r = .27$). An exploratory factor analysis found that a two-factor solution accounted for 43.41% of total variance, indicating a clear separation of items between one factor named “congruence/ experiential fluidity” (consisting of all positively worded items), and a second factor named “incongruence/ experiential constriction” (containing all negatively worded items). Evidence of convergence and discrimination between scores on the SI and other selected measures indicated strong positive associations with self-esteem and accepting one’s own emotions, and strong negative correlations with experiencing a lack of emotions and psychological distress.

Using these results, Freire (2007) produced a 31-item version (SI-31) and tested it using data collected from two USA-based student samples and a new UK-based sample of trainee and qualified counselors. The SI-31 scores demonstrated adequate temporal consistency (*r* = .66) and replicated the original findings: internal consistency remained high (α = .94), and a two-factor solution (44.67% of total variance) reflected the direction of item wording. Freire tested the construct validity of this two-factor model with a range of instruments, including the CORE-Outcome Measure (CORE-OM; Barkham et al., 2015) and the NEO Five Factor Inventory (McCrae & Costa, 2013). For Factor 1 (positively worded items), the results indicated moderate negative correlations with psychological distress (*r* = -.59; Factor 2, *r* = .56) and neuroticism (*r* = -.45; Factor 2, *r* = .57), moderate positive correlations with extroversion (*r* = .48; Factor 2, *r* = -.20), agreeableness (*r* = .46; Factor 2, *r* = -.31), and conscientiousness (*r* = .42; Factor 2, *r* = -.23), and a small positive correlation with openness (*r* =.28; Factor 2, *r* = -.04). When tested with a one-factor model (all SI items), these correlations were maintained, increasing slightly for psychological distress (*r* = -.66), neuroticism (*r* = -.60), and agreeableness (*r* =.46), and decreasing slightly for extroversion (*r* = .48), conscientiousness (*r* = .42), and openness (*r* = .22). Indeed, Freire found a moderate negative correlation (*r* = -.46) between the two proposed factors and acknowledged the possibility that they may represent an artefact of using positively and negatively worded items (e.g. Solís Salazar, 2015).

Zech et al. (2018) reported the psychometric properties of a French-language version of the SI using data collected from Belgian student and patient samples. Similar to Freire (2007), Zech et al. identified very good inter-item consistency, adequate temporal consistency, and evidence of convergence and discrimination consistent with predictions: a high positive correlation with emotional intelligence; moderate positive correlations with extraversion and agreeableness; moderate negative correlations with indicators of alexithymia and neuroticism (in students), along with symptoms of anxiety and depression (in patients). A principal components analysis indicated a clear separation of positively and negatively worded items with a moderate negative correlation (r = -.45) between the two factors. Zech et al. proposed "a hybrid model of an over-arching Congruence-Incongruence dimension, with two sub-factors" (p. 176).

Aims of this Study

The main goal of this study was to examine the validity of the SI for use as an outcome measure in counseling using data archived by a UK-based university counseling research clinic. We had collected data using the SI-31 in the research clinic since 2007. In 2012, Elliott & Rodgers used inter-item correlations and exploratory factor analyses to assess the internal consistency of the SI-31 data collected. This initial attempt to validate the SI-31 indicated potential item redundancy, resulting in the introduction of a 16-item version (SI-16; Table1) for use in the research clinic. We planned to conduct a robust evaluation of the validity of the SI-16 using the data collected. In addition, we wanted to address the unresolved question of the SI's dimensionality.

We proposed to conduct a two-part investigation: in Study 1, to test the SI's internal validity (i.e., internal structure, dimensionality) using Rasch modeling (Bond & Fox, 2015); in Study 2, to scrutinize its external test validity (i.e., sensitivity to change) for evidence of its applicability as an outcome instrument. An unexpected product of our study was the development of a new shorter version of the instrument that maintained fit to the Rasch model and did not result in construct underrepresentation (i.e., the removal of vital aspects of what it means to be fully functioning; Spurgeon, 2017). We present the research questions that guided each study at the beginning of the relevant section.

Study 1

Research Questions

1. Can participants distinguish among the five categories in the SI rating scale?
2. Does the SI measure a unidimensional concept?
3. Is the SI able to distinguish meaningful levels of person ability among participants?
4. Do measurement gaps exist in the SI hierarchy of item difficulty that may indicate construct underrepresentation?

Method

There has been an emerging trend in the last two decades to use Rasch modeling for the development and validation of tests used in the counseling field (e.g., Saks et al., 2020). Rasch modeling rejects the assumptions within classic test theory that all items are equally weighted and represent interval-level scales. Instead, Rasch modeling transforms raw scores into standardized units of measurement (logits) and uses unidimensional linear models to detect problems in both dichotomous and polytomous rating scales, and calculate probable or expected scores based on parameters that represent "person ability" and "item difficulty". If the expected measures (i.e., mean and standard deviation) for person ability and item difficulty cover the same area in the linear model then an instrument is well-targeted to the sample that it is being used to assess. Rasch modeling enables examination of the internal structure of the construct, for example, by using expected scores to assess dimensionality and to order items according to difficulty. Furthermore, Rasch modeling evaluates individual items based on misfit to expected scores using infit and outfit measures: infit is a weighted residual placing more emphasis on unexpected responses close to the item's mean measure; outfit is unweighted, produced by averaging the residual variance in scores across items and can be influenced by unexpected responses at the extreme ends of the scale. These analyses produce evidence that can guide a thoughtful theory-driven approach to measure development rather than the routine application of standardized criteria (Bond & Fox, 2015).

Analysis

We chose Rasch modeling for our first study because it offered the means to test for evidence of the validity of the SI as an outcome instrument in three key areas: response processes, internal structure, and test content (American Educational Research Association et al., 2014). First, we planned to test the utility of the rating scale by analyzing the functioning of the five-category rating scale. This can be done by examining category use statistics (i.e., category frequencies, mean measures, threshold calibrations; see Bond & Fox, 2015, pp. 248-252), seeking response categories that are disordered (i.e., not increasing monotonically) or misfitting the Rasch model, and identifying the optimal number of response categories. Second, we intended to investigate the internal structure of the SI using standardized fit statistics, item misfit statistics, a Dimensionality map, and a Construct KeyMap (see pp. 130-134). This series of steps enabled us to systematically test and accrue evidence of the internal structure of the instrument. Third, we proposed to inspect test content by examining the order of item difficulty produced by the Construct KeyMap and assessing fit with the concept that the SI purports to measure: the fully functioning person.

We used Winsteps (Linacre, 2017) to analyze our SI data according to Andrich's (1978) rating scale model, and Bond and Fox (2015) to support our interpretation of the results. In this article, we present the application of the Rasch model in our study; information about the technical aspects of polytomous Rasch models can be found in Bond and Fox (2015; in particular, appendix B).

Participants

The 385 participants were clients of a counseling research clinic situated in a UK-based university between 2007 and 2016 who consented to take part in research activities alongside the counseling process. The sample contained data from two protocols: a general 'practice-based' (PB) protocol (*N* = 294), which offered up to 40 sessions of person-centered counseling (PCT; Mearns et al., 2013) with the possibility of extension if mutually agreed;

and a specialist protocol (*N* = 91) offering 20 sessions of either PCT or emotion-focused counseling (Elliott & Greenberg, 2021) to people experiencing social anxiety difficulties (SA). The majority of participants in both protocols were female (PB – 65.3%; SA – 56%) and White European (PB – 96.7%; SA – 94.1%) and of a similar mean age (PB – 35.9 years, range 18-73; SA – 33.2 years, range 18-60). The two research protocols were approved by local National Health Service and university ethics committees.

Instrument: The Strathclyde Inventory (SI)

The Strathclyde Inventory (SI-16; Table 1) is a 16-item self-report instrument with a five-category Likert-type scale designed to measure Rogers' concept of the fully functioning person. The SI-16 is a shortened version of the SI-31 (Freire, 2007) produced by Elliott and Rodgers (2012) following an analysis of the internal consistency of the SI-31 using data collected from clients in the research clinic from 2007 until 2012. Decisions for the removal of these items were guided by low item-total correlations (< .5), and a pragmatic approach to content validity (i.e., perceived overlap in meaning of items). Internal consistency was maintained (SI-16, α = .93; SI-31, α = .95) when recalculated using the same dataset. An exploratory factor analysis of SI-16 found that two factors (positively worded items, and negatively worded items) could explain 50.8% of the total variance in the data. When completing the SI , participants are asked to read each statement, consider how often it has been true for them during the last month, then mark the box that is closest to their experience using a five-category rating scale (scored 0, 1, 2, 3, 4) with the following anchor words: never, only occasionally, sometimes, often, all or most of the time.

Procedure

Participants completed the SI at designated time points: before counseling began, at regular intervals during counseling, at the end of counseling, and at two optional follow-up points (six and 18 months after the end of counseling). Thus, the number of data collection points at which the SI was completed per participant varied (range = 1-10; M = 2.89; SD = 1.72; median = 3; mode = 1) according to the duration of their counseling and participation in follow-up procedures. In order to maximize the potential dataset for our study, we decided to include 652 observations collected using Freire's (2007) 31-item version (SI-31; PB = 436, SA = 216), extracting the data collected on the 16 items that had subsequently formed the SI-16 (Table 1), with the 522 observations (PB = 410; SA = 112) collected using the SI-16.

As a dataset containing multiple observations obtained from the same participants was likely to violate the key statistical assumption of independence of observations, we created a sub-sample that included only one observation per participant. We considered applying the conventional approach of using pre-therapy observations only. However, as the SI is intended for use as an outcome instrument and therefore administered to the same participant on multiple occasions, we decided that it would be more appropriate for ecological validity to select observations across the range of data collection points. In addition, we expected this would provide a wider representation of the construct being measured. We used an online true random number service (random.org) to select one observation from each participant who had completed the instrument on more than one occasion. This process resulted in an 'independent' dataset of 385 observations. Table 2 presents the original full dataset and the subsequent "independent" dataset, identifying number of observations from each protocol, SI version, and data collection point.

Results

Can Participants Distinguish between the Five Response Categories in the SI Rating Scale?

It is fairly typical to find that participants have had difficulties in using a five-category rating scale, perhaps due to the 'middle category measurement flaw' illustrated by Bradley et al. (2015). Table 3 shows the total frequency (count), mean measure, and the infit and outfit mean squares (MNSQ) for each category, all of which fall well within the range (0.6 – 1.4) recommended by Bond and Fox (2015, p. 273) for scores obtained from rating scales. The category that contained most random responses (infit MNSQ = 1.28; outfit MNSQ = 1.31) was 4, "all or most of the time" (or "never" for items that were reverse scored), indicating approximately 30% more randomness than predicted by the model. This was also the category used least frequently, consistent with its position as an extreme category within the five-category scale.

Table 3 also reveals the threshold calibrations between response categories increased monotonically across the measure. This indicates that, on average, persons with increasing functioning endorsed progressively higher categories. Linacre (1999, p. 119) recommended that there should be at least 1.0 logit between thresholds in a five-category scale, which was the case here. The ordered nature of the SI-16 rating scale can be seen clearly in the form of a probability curve (Figure 1), which shows distinct thresholds between categories. Indeed, this probability curve is unusually clean and clear, indicating that participants were able to distinguish between the five categories on the rating scale, leading us to the conclusion that the SI-16's five-category rating scale was functioning as intended.

Does the SI Measure a Unidimensional Concept?

The starting point for assessing fit to the Rasch model is an examination of the standardized residuals for persons and items, known as the 'fit statistics'. Table 3 presents the SI-16 fit statistics: first for persons (*N* = 385), and then for items (*N* = 16). For persons, the mean raw score was 30.3 (*SD* = 11.7), demonstrating a substantial variation in raw scores between persons. The mean measure (-.11) represents the mean ability of all persons in the sample. A perfect fit between person ability and item difficulty according to the Rasch model would be zero, therefore this result indicates a close match between participants' ability and the instrument's difficulty, suggesting that the SI is well-targeted for this sample of clients.

This evidence of good fit is supported by the closely matching infit and outfit MNSQ statistics for persons (1.01) and items (infit 1.00; outfit 1.01).

We examined the misfit statistics for each individual item to identify underfit and overfit at item-level (see Supplemental Table S1). The outfit MNSQ for all individual items ranged from 1.51 (item 7) to .61 (item 15), within the range (0.5 – 1.5) identified as productive of measurement by Linacre (2017). The point-measure correlation, indicating how well the individual item aligns with the instrument as a whole, ranged from .54 (item 7) to .77 (items 14 & 15). The standardized *Z* scores reflected the MNSQ ordering of items but contained some unexpectedly large results: nine items had infit and outfit *Z* scores that exceeded +/-2.0. Three of these (items 7, 6 & 5) had positive *Z* scores, indicating underfit: unexpected responses, possibly the result of guessing. The negative *Z* scores reported for the remaining 6 items (items 15, 14, 4, 12, 8 & 16) suggested overfit, likely due to item dependence. However, we noted that these results may reflect sample size: Smith et al. (2008) identified that standardized fit statistics for polytomous Rasch models are more likely than mean square fit statistics to be sample-size dependent, picking up small misfits and amplifying them.

Next, we investigated the dimensionality of the measured construct by conducting a Rasch principal components analysis (PCA) to produce a Dimensionality map (Bond & Fox, 2015). Rasch PCA examines the standardized residuals that remain after the linear Rasch model has been extracted, looking for indications of any other common variance within the residuals. The total variance found in our data measured 40.8 eigenvalue units; of this, 24.8 units (60.7%) was explained by the measure, much greater than the typical value (40% - 50%) that supports the presumption of unidimensionality, according to Linacre (2017).

A central feature of Rasch PCA is the ability to contrast the extracted residuals. Residual variance should be random and without structure. In our analysis, five contrasts were extracted, but only the first contrast was greater than 2.0 eigenvalues, the recommended threshold for random variance warranting further investigation (Linacre, 2017). The loading of items onto the first contrast clearly revealed the now familiar pattern of one cluster of positively loaded items (the six reverse scored items) and one cluster of negatively loaded items (the ten other items). As recommended by Linacre, we calculated mean person measures for the two item groups, first, positively loaded items, then, negatively loaded items, and correlated the results ($r = .64$). This was far larger than the correlations detected by Freire (2007) and Zech et al. (2018), providing stronger evidence that these groups of items are not measuring something substantively different.

Is the SI Able to Distinguish Meaningful Levels of Person Ability among Participants?

The person separation index (2.92; see Supplemental Table S2) can be converted into a strata index (see Bond & Fox, 2015, formula 16). A strata index represents the number of measurably distinct strata (i.e., layers or levels) of person ability (or item difficulty) that can be supported by the data, an important requirement for an outcome measure. This calculation for persons was 3.6, indicating that the SI-16 can distinguish at least three distinct levels of ability in the data. The item separation index of 6.48 converted to 8.3 strata, indicating at least eight steps in the hierarchy of item difficulty. These item and person strata are visible in the Construct KeyMap (Figure 2; Bond & Fox, 2015), an “expected score” item-person matrix showing the average category rating predicted for each item according to person ability.

In Figure 2, the average person measure is almost in line with the mid-point (-.11) of the x-axis, indicating an overall good match between the sample and the instrument. The three levels of the construct proposed by the person strata index could be represented by low functioning (the zone more than one standard deviation below the mean), high functioning (the zone more than one standard deviation above the mean), and average functioning (the zone that is within one standard deviation on either side of the mean).

The Construct KeyMap orders items from easiest to endorse, therefore within the capability of those with a lower level of functioning (item 10 – *I have made choices based on my own internal sense of what is right*), to most difficult to endorse, therefore requiring greatest functioning (item 12 – *I have lived fully in each new moment*). We reviewed the order of items to assess if this made sense from a theoretical perspective. We considered that it did, noting that the content of items represented a gradual overlapping process of increasing self-awareness, self-trust, self-acceptance, openness to self, and openness to others; an accumulative process consistent with the fully functioning person construct. However, we noted that item 10 did not fit comfortably as a first step, indicating potential construct underrepresentation. We expected that someone with a lower level of functioning would struggle to know their own sense of what is right and act upon it; something else would have to occur first (e.g., increased awareness of what they felt and believed and a growing willingness to trust in themselves). This suggested gaps at the lower end of the instrument.

What Measurement Gaps Exist along the SI Hierarchy of Items that May Indicate Construct Underrepresentation?

As an instrument for use within a counseling setting, the SI must be able to measure lower levels of functioning. Therefore, we reviewed items discarded from the SI-31 when the SI-16 was created, and selected four items that described experiences of relevance for persons low in functioning. These items were inserted into a new version (SI-20) in the same positions they had held in SI-31. The four new items (with their SI-20 numbering) were: 12 - *I have felt myself doing things that were out of my control* (R); 14 - *I have been aware of my feelings*; 18 - *I have felt myself doing things that are out of character for me* (R); and 19 - *I have accepted my feelings*.

Next, we extracted data collected on these 20 items from the SI-31 subsample within the independent dataset (*n* = 216; Table 2) and repeated the series of Rasch analyses. Supplemental Table S2 presents a comparison of the fit statistics for SI-16 and SI-20, indicating a very similar, indeed slightly better, fit to the Rasch model. An examination of item misfit statistics for the SI-20 (see Supplemental Table S3) identified some minor shifts in the order of items and a slight reduction in *Z* scores, probably due to the smaller sample size (Smith et al., 2008).

The statistics for the four new items demonstrated a comfortable fit within the parameters set by the existing group of items. Item 12 was found to be the second most underfitting item of the group of 20, but still within acceptable limits. Items 18 and 14 were better fits. Item 19 was somewhat overfitting but not extreme within the overall group. The mean measures of three items (18, 14 and 12) indicated a better fit for persons with lower functioning. To confirm this observation, we produced a Construct KeyMap (see Supplemental Figure S1). As expected, these three items clustered at the lower end, below and just above the original lowest difficulty item. This positioning made sense theoretically, offering additional opportunities to capture the early development of persons at the lower end of the scale and improving the overall validity of the instrument. However, there was now less differentiation in the middle section, suggesting item redundancy. We decided to proceed with Study 2, then evaluate the evidence that we had gathered, and produce a briefer version.

Study 2

Research Questions

1. Are SI scores temporally consistent prior to the start of counseling? (i.e., test-retest reliability; also required for calculating reliable change)

2. Do scores on the SI change over the course of counseling? How does change in scores on the SI compare with change in scores on general and individualized measures of distress? (i.e., sensitivity to change; discriminant validity)
3. Are some items within the SI more sensitive to change? Do any SI items change in meaning for participants by the end of counseling? (i.e., sensitivity and stability of meaning of individual items; required for further development of instrument)
4. According to their SI scores, what percentage of participants recovered, improved, or deteriorated by the end of counseling? (i.e., sensitivity to "clinically significant change"; Jacobson & Truax, 1991)

Method

To investigate the external validity of the SI for use as an outcome measure, we returned to our original dataset but made the decision to include only the observations collected from clients who participated in the Practice-Based (PB) protocol. Diverse in their reasons for accessing counseling, it was likely that this sub-sample would reflect a typical heterogenous client population accessing counseling in routine practice. Data collection took place as part of a single protocol organized on the principles of an open clinical trial with repeated measures across treatment.

Participants

We included all participants (*N* = 225) who had completed the SI before commencing and at least once prior to ending counseling. If a participant had left counseling without taking part in post-counseling procedures, or were still in counseling, then we adopted a last observation carried forward approach in which their last SI score was treated as their post-counseling score. The majority of participants were female (144, 64%; male = 80, 35.6%; non-binary = 1, 0.4%) and White European (*N* = 215, 95.5%; Asian Indian/Pakistani = 4, 1.8%; Other = 6, 2.7%) with an age range of 18-67 years (*M* = 35.53, *SD* = 11.7). Participants were offered up to 40 sessions of counseling but this maximum could be exceeded if mutually agreed. The participants in this dataset attended 3-70 sessions of counseling (*M* = 25.21; *SD* = 13.86; median = 23; mode = 40). All participants consented to take part in research activities alongside the counseling process.

Data Collection

The SI was administered to participants by their researcher (a member of the research clinic team who was not their counselor) at each data collection point, along with CORE-OM (Barkham et al., 2015) and the simplified Personal Questionnaire (PQ; Elliott et al., 2016). These instruments formed the standard set of outcome measures for the PB protocol. The SI was presented to participants within this battery of outcome measures, usually the second or third of the three instruments completed on each occasion.

CORE-OM. The CORE-OM is a 34-item self-report instrument with a five-category Likert-type scale designed as a general measure of distress experienced in the past seven days. The internal consistency (α = .91 - .95) and test-retest reliability (.88) of CORE-OM scores is excellent (Barkham et al., 2015). A clinical significance cut-off score of 1.0 and approximate RCI minimum value ($p < .05$) of .5 is recommended (Barkham et al., 2015). Skre et al. (2013) confirmed earlier studies that found the CORE-OM is best scored as two scales: one, general psychological distress (28 items), and the other, risk (6 items). The internal consistency of scores on the CORE-OM and its two scales in our dataset was consistent with the populations in previous studies: slightly higher for CORE-OM (34 items; α = .95) and the psychological distress scale (28 items; α = .95), and somewhat lower for the risk scale (6 items; α = .69).

Personal Questionnaire. The PQ is a client-generated outcome measure in which participants identify specific difficulties they wish to address in counseling and use a seven-category rating scale to indicate how much each problem had bothered them during the past seven days. Elliott et al. (2016) analyzed PQ data from five samples, including participants in this dataset, and reported good internal consistency both between-clients and within-client (α = .70-.80); consistent temporal reliability (*r* = .57); strong correlations with a selection of standardized outcome measures (typically .30-.60); and large pre-post effect sizes (*d* = 0.8–1.7). They recommended a clinical significance cut-off score of 3.25 and RCI minimum value of 1.5 (*p* < .05).

Data analysis

First, we checked our data for normality of distribution and outliers, and found no issues of concern. Then, we addressed our research questions using standard analyses and statistics (e.g., Pearson's correlations, paired samples t-tests, Cohen *d*, Jacobson & Truax criterion C) to assess temporal consistency (question 1), sensitivity to change (questions 2 & 3), and clinical significance (question 4). In addition, we used Differential Item Functioning (Bond & Fox, 2015) to seek evidence of change in item difficulty over the course of counseling (question 3).

Results

Are SI Scores Temporally Consistent Prior to the Start of Counseling?

We expected little change in a participant's functioning in the immediate period prior to beginning counseling. In our dataset, 44 participants had completed the SI on two occasions before commencing counseling: the first at their intake interview, the second immediately before their first counseling session. The median duration between the two time points was 15 days (*M* = 44 days; range = 2 – 144 days). Using Pearson's correlation, we found adequate test-retest reliability between scores collected at the two time points (*r* = .81), providing evidence that these clients' SI scores were temporally consistent prior to commencing counseling.

Do Scores on the SI Change over the Course of Counseling?

We predicted that pre-post change in SI scores would be large and statistically significant but that, as it sought to measure a form of experiential functioning rather than symptoms of distress, then the size of this change would be somewhat smaller than for the PQ and the CORE-OM sub-scales (Elliott, 2001). Table 4 presents the results of our analyses. We excluded listwise 20 cases from our dataset that were missing either PQ or CORE-OM data.

Is the Change in Scores Statistically Significant? On Average, what Size is the Change? Does this Differ across Versions? The mean SI score at pre-counseling was 1.80 (*SD* = 0.66); at the end of counseling it was 2.48 (*SD* = 0.79). Higher scores indicate improvement. Using a paired-samples t-test, the difference (0.68) was statistically significant (*t*(204) = -13.29, *p* <.001, *r* = .50); a large effect (*d* = .93; CI95 = .73, 1.14), according to Cohen (1988; *d* = .2, small; .5, medium; and .8, large effect). In order to compare sensitivity to change across SI versions, we conducted paired-samples t-tests using the data from participants who completed the same versions (either SI-31 or SI-16) at pre- and post-counseling: Change in pre-post mean scores recorded was statistically significant and large for each version: SI-31, *n* = 90, *t*(89) = 8.35, *p* < .001, *r* = .48, *d* = .91, CI95 = .60, 1.21; SI-16, *n* = 90, *t*(89) = 9.68, *p* < .001, *r* = .44, *d* = 1.08, CI95 = .77, 1,39. We did not test the difference between the two correlations because it was clearly so small (.04). Similarly, the difference in effect size (*d*) between the two versions was only .17, a small effect, and unlikely to be statistically significant.

How does Change in Scores on the SI Compare with Change in Scores on the CORE-OM and PQ? We repeated these analyses for the pre- and post-counseling data collected from the same 205 participants using the CORE-OM subscales and the PQ (Table 4). For the CORE-Distress subscale the difference (.74) was statistically significant and represented a large effect (*d* = .95; CI95 = .75, 1.16), almost identical to the SI. This finding is interesting because it indicated that the two measures detected a very similar sensitivity to change in participants. For the PQ we found a statistically significant difference representing a large effect (*d* = 1.55; CI95 = 1.33, 1.77), consistent with Elliott's (2001) findings for individualized outcome instruments.

Next, we used two-tailed Pearson correlations to investigate the relationship between pre-post change in scores recorded for each of the three instruments. The results (Table 4) confirmed Freire's (2007) findings: we found a strong association between pre-post change on the SI and CORE-Distress (*r* = .75, CI95 = .68, .80), and a more moderate relationship between the SI and PQ (*r* = .54, CI95 = .44, .63). A greater distinction between the SI and CORE-Distress was indicated by correlations with the CORE-Risk subscale: with CORE-Distress (*r* = .47, CI95 = .36, .57); with SI (*r* = .25, CI95 = .12, .37). This finding suggested that the amount of change in functioning has only a small association with the amount of change in level of risk during counseling; in other words, that the SI is measuring something distinctively different.

Are Some Items within the SI More Sensitive to Change? Do Any SI Items Change in Meaning for Clients by the End of Counseling?

We calculated the statistical significance and effect size of any difference between mean pre- and post-counseling scores for each item in SI-20 (see Supplemental Table S4). The pre-post difference in scores was statistically significant (*p* < .001) for all items except item 10 (*I have made choices based on my own internal sense of what is right*), which was statistically significant but at a lower standard (*p* < .01), and item 14 (*I have been aware of my feelings*). Five items demonstrated a large pre-post effect (*d* > .75): items 2, 7, 9, 13, and 17. This result suggested that these items were the most sensitive to the type of change that participants experienced over the course of counseling. Two items had small effects (*d* < .44): items 10 and 14.

We investigated whether the *meaning* of any item had changed for participants between the start and end of counseling. For this analysis, we returned to Rasch modeling. Differential item functioning (DIF) compares individual item difficulty between groups. In this case, we used DIF to analyze the difference in item difficulty for participants when completing the SI at pre-counseling compared with post-counseling. The DIF identified three items with a statistically significant difference: item 5 ($t = 2.84$, $p < .01$); item 10 ($t = 4.53$, $p < .0001$); and item 14 ($t = 3.34$, $p < .01$). However, only two of these items had a DIF contrast greater than .5 logits, the minimum size recommended by Bond and Fox (2015) to merit further investigation: item 10 = .53; item 14 = .59. In both cases, we found that the items were more difficult (i.e., the mean score was lower) at post-counseling. This result suggested a change in participants' perception of each item. This could be that their interpretation of the meaning of the item had changed (e.g., for item 10, what it *really* means to make choices based on an internal sense of what is right), or alternatively that their understanding of its relevance to their own experience changed (e.g., they have developed a greater appreciation of the challenge for them in making decisions based on an internal sense of what is right). Both of these possible explanations suggest some degree of increased self-awareness, a paradoxical indicator of increased functioning.

According to Their SI Scores, What Percentage of Participants Recovered, Improved or Deteriorated by the End of Counseling?

First, we carried out a small meta-analysis that combined the results of this study with those of previous studies (Folkes-Skinner, 2011; Freire, 2007; Zech et al., 2018), to calculate a standardized clinical significance cut-off score and reliable change indices based on SI data collected from clinical and non-clinical populations (RCI; Jacobson & Truax, 1991, criterion C): for more information, see Supplemental Note S1 and Supplemental Table S5. This resulted in a clinical significance cut-off score of 2.36, and minimum RCI values of .97 (*p* <

.05) and .64 ($p < .20$). Next, we used these indices to calculate reliable change (improvement or deterioration) and clinically significant change (reliable change plus movement from the clinical range across the clinical significance cut-off score, indicating recovery) for each individual participant in our dataset.

Applying the larger RCI value ($p < .05$) to our data, 31.1% of participants reliably improved (26.2% recovered, 4.9% improved but not recovered), and 0.9% deteriorated (see Supplemental Table S6). The largest group of participants (68%) were those whose scores did not change sufficiently to meet the criterion for reliable change. Almost one fifth (20.4%) of participants commenced counseling with scores in the non-clinical range, a ceiling effect that made it difficult to register reliable change.

Development of SI-12

The evidence collected in our first study indicated that removal of items was desirable due to a density of similar items in the middle range of the instrument, We were also motivated to create a briefer version of the SI as brief instruments are considered to be less onerous for participants (e.g., Rolstad et al., 2011). Our results suggested that items could be removed carefully without significantly harming the instrument's fit to the Rasch model nor introducing construct underrepresentation (Spurgeon, 2017).

First, we gathered the results of the various item-level analyses conducted in our two studies (see Supplemental Table S7). In addition, we conducted a reliability analysis of the data collected on the SI-20 items. This confirmed that all corrected item-total correlations and squared multiple correlations were within acceptable limits, and that no items, if removed, would increase the Cronbach's alpha. There were no inter-item correlations greater than .7, which would have indicated a high degree of correlation, but we identified eleven pairs of items with inter-item correlations greater than .6 (see Supplemental Table S8), and used this information as a confirmatory check of items that we identified as potentially redundant. An outline of our process and decisions in reviewing items for removal is presented in Supplemental Note S2.

Testing Alternative Brief Versions.

Developing instruments using Rasch modeling is an iterative process of creating and comparing alternative possibilities. First, we applied the results of our item review in stages, creating three alternative brief versions: a 14-item version, removing items 5, 6, 7, 16, 17, 19; a 13-item version, removing item 8; and a 12-item version, removing item 10. Next, we compared the fit statistics and variance explained by the measure for SI-20 and each alternative brief version and noted a steady decrease in the person reliability index, separation, and strata as the number of items was reduced (see Supplemental Table S9). The item reliability index remained constant at .98, providing reassurance that these groups of items were likely to perform in the same way with another sample of participants of similar ability (Bond & Fox, 2015). Item separation and strata increased as the number of items decreased, reflecting an increasing differentiation in item difficulty. Variance explained by the measure improved in briefer versions (SI-12 = 63%) as did item misfit (see Supplemental Table S10).

These statistics confirmed that it was possible to reduce the items contained in the SI and not harm (indeed, slightly improve) its overall fit to the Rasch model. The SI-12 was the most attractive version to adopt because it was the briefest. We produced a Construct KeyMap (Figure 3) to make two final checks: first, that the items were sufficiently distributed across the instrument to fit persons across the range of ability; and second, that the content of the remaining items continued to make sense theoretically, without any obvious loss of meaning. As Figure 3 demonstrates, the mean measures of 90% of participants fell within two standard deviations above and below the mean (highlighted by the two vertical dot-dash lines), indicating a good match with the SI-12. Finally, we reviewed item content and noted that the remaining items retained the six layers of development originally noted in the SI-16 and expanded in the SI-20. We concluded these items represented a credible description of the fully functioning construct, consistent with Rogers' writings. As a result, we made the decision to accept SI-12 (Figure 4).

Discussion

Five key findings from this study are particularly relevant for counselors interested in using the SI as a measure of outcome in their counseling practice. First, we demonstrated that the five-category rating scale functioned well. As an instrument intended for use in counseling settings, it is essential that participants can distinguish and select the rating scale category that best fits their personal experience. Although five-category rating scales can often cause difficulties (e.g., Bradley et al., 2015), our analysis did not find evidence of problems is the use of rating scale categories for participants in this study.

Second, we found evidence that the SI can be applied as a unidimensional instrument acceptable to the stringent Rasch model. This finding provides an answer to a long-standing question and is consistent with Rogers' proposition that the process of constructive therapeutic change exists on a continuum (Rogers, 1961/1967). It means that counselors can feel confident in calculating a total mean score when using the SI as a representation of their client's functioning over the course of counseling.

Third, the effect size of pre-post change in mean SI scores for participants in our second study was large (.93), although we found a more modest result when analyzing individual scores for clinically significant change (Jacobson & Truax, 1991; criterion C). According to Lambert (2013, p. 178), it is typical to find that the effect size statistic overestimates the proportion of individuals whose change in scores can be defined as clinically meaningful. The results of this study confirmed that the SI demonstrates a sensitivity to change in scores over the course of counseling that is equivalent to other commonly used outcome measures, and indeed that the SI may be capturing a higher degree of change than is typical for measures of experiential functioning (Elliott, 2001).

Fourth, the SI is measuring something different yet related to the type of distress captured by the PQ and CORE-OM. Rogers' theory of change, while not prioritizing the reduction of distress, clearly acknowledges a relationship between increased functioning and decreased distress: "The feeling of reduction of inner tension is something that clients experience as they make progress in 'being the real me' or in developing a 'new feeling about myself'" (Rogers, 1951, p. 513). More recently, Warner (2017, p. 109) agreed, stating "such processing and self-cohesion allow the development of a 'congruent' version of self that resonates with the person's whole-body experience and minimizes psychological symptoms." As proposed by Levitt et al. (2005), counselors who work with Rogers' theory of change, or a similar model of psychological growth, may choose to use the SI as an alternative to outcome instruments focused on symptoms of distress.

Fifth, in Study 1 we demonstrated that it is possible to use Rasch modeling to distinguish meaningful levels, or degrees, in the process of becoming more fully functioning. Indeed, perhaps our most interesting finding is this hierarchical relationship among SI items, indicating a potential sequence in the process of becoming more fully functioning that dialogs with the characteristics identified by Rogers (1959, 1961/1967). First, self-trust (internal locus of evaluation) mediates self-awareness (openness to experience), enabling the accurate symbolization of experiences into awareness; second, self-trust deepens into self-acceptance (trust in one's organism, unconditional self-regard) facilitating the flexible integration of new experiences into a coherent whole-self construct; and third, self-acceptance becomes the basis for openness to self (willingness to be a process) and openness to others (living in harmony with others) highlighting the relational nature of the fully functioning process. Indeed, openness to others may be understood as an *effect* of increasing openness to self; the pivot point at which becoming more fully functioning shifts from being an intrapersonal process to an interpersonal experience. It demonstrates that we may only become truly open to others (i.e., lower our guard, remove our mask) when we trust, accept and are able to be open to ourselves. As Rogers outlined in the 18th of his Nineteen Propositions:

> When the individual perceives and accepts into one consistent and integrated system all his sensory and visceral experiences, then he is necessarily more understanding of others and is more accepting of others as separate individuals. (1951, p. 520)

This finding is also consistent with the more recent work of Stevens (2017) who found evidence of authenticity as a mediator between attachment style and affective functioning, and proposed that "if individuals cannot be genuine with themselves, then genuine behaviors and genuine relationships will be hard to establish" (p. 408). According to their SI scores, the struggle to trust and be open with others remained a concern, to some degree, for many of our participants, even at the end of counseling. This finding cautions counselors inclined to assume that client trust and openness in the therapeutic relationship is something that can be easily won before the "real" work begins.

Implications for Counseling Practice

In summary, the SI offers clients an opportunity to reflect on their experiences from an alternative perspective to symptom-oriented instruments, one that is more clearly aligned with the growth-oriented attitudes, values, and language of counseling practice. We believe that this shift in emphasis is an important one as it strengthens the message that we give our clients about their potential for growth, characterized by Rogers as the fully functioning person, offered by the counseling process. On the basis of the results reported in this article, we invite counselors to adopt the SI-12 within their practice: asking clients to complete the instrument before counseling begins and, at minimum, once more at the end of their counseling process; scoring each observation by calculating a single mean score; and interpreting change in scores by calculating the difference between mean scores, then comparing this with the clinical significance cut-off score of 2.36, and minimum RCI values of .97 (p < .05) or .64 (p < .20) reported here.

We encourage counseling practitioners to use their clients' SI scores to reflect with their clients, and also their supervisors, on changes in their clients' processing during counseling. Counselor educators may consider using the SI as a growth-oriented measure of the development of counselors in training for the purpose of self-reflection, course evaluation, and research. The use of the SI in the training context has been reported by Folkes-Skinner (2011), and Brison et al. (2015).

Limitations and Recommendations for Future Research

The generalizability of our results is limited by the cultural context in which the data was collected: provided by UK-based clients accessing free counseling services situated within a university-based research environment. Each of these characteristics have potential implications, including assumptions and norms implicit in UK culture and the manifold socio-economic issues that may attract clients to access free counseling, but also deter them from using a service based on a university campus, and require sufficient literacy to take part in research activities.

Future research should continue to explore the construct measured by the SI, for example by Rasch modeling the item-level relationship between SI items and measures of associated variables such as the NEO-PI-3 (McCrae & Costa, 2013), and the Authenticity Scale (Bayliss-Conway et al., 2020). Indeed, many questions remain. What can we learn about the change processes that occurred in counseling for those clients whose scores on the SI indicated reliable improvement (or deterioration)? What happened over the course of counseling for the 68% of participants in this study whose scores did not change by the minimum value required to indicate reliable change, and the 20.4% of participants who began counseling with SI scores in the non-clinical range on the measure? Our results raise interesting questions about how participants perceived themselves in relation to SI items, especially at the beginning of the counseling process. If a client is relatively incongruent at pre-counseling, therefore low in self-awareness, then it is possible that they may report a higher (i.e., better) than expected score. It may be that as their self-awareness increases in counseling, clients become more aware of their limited functioning, resulting in lower subsequent scores on the instrument. This kind of 'worse before better' pattern was identified by Owen et al. (2015) as one of three distinct trajectories that can occur across longer periods of counseling, and may represent a 'response shift', identified as a potential issue when using self-report instruments in counseling outcome research (e.g., Murray et al., 2018). We recommend that future research using data collected on the SI should seek to explore these questions.

Conclusions

We found evidence of excellent internal consistency, clearly replicating the results of previous studies using SI data collected from clinical and non-clinical populations. Going beyond previous studies, Rasch modeling indicated that the instrument is unidimensional and that clients in counseling can discriminate between the five response categories offered to them. We obtained evidence that the SI demonstrates temporal consistency at pre-counseling and is sensitive to change in scores over the course of counseling, an important requirement for an instrument designed to measure the outcome of counseling. Finally, we used findings from our two studies to inform the creation of a briefer, more user-friendly 12-item version that maintained fit to the Rasch model and construct representation.

[Insert url for Online Supplemental Material here]

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* American Psychological Association.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*(4)*,* 357-374. https://doi.org/10.1007/BF02293814

Barkham, M., Mellor-Clark, J., & Stiles, W. B. (2015). A CORE approach to progress monitoring and feedback: Enhancing evidence and improving practice. *Psychotherapy, 52*(4), 402-411. http://doi.org/10.1037/pst0000030

Bayliss-Conway, C., Price, S., Murphy, D., & Joseph, S. (2020). Client-centred therapeutic relationship conditions and authenticity: A prospective study. *British Journal of Counselling & Guidance.* https://doi.org/10.1080/03069885.2020.1755952

Bayne, H. B., & Hankey, M. S. (2020). Development and Rasch analysis of the Empathic Counselor Response Scale. *Measurement and Evaluation in Counseling and Development, 53*(3), 182-194. https://doi.org/10.1080/07481756.2019.1691462

Bond, T. G. & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences*. Routledge.

Bradley, K. D., Peabody, M. R., Akers, K. S., & Knutsen, N. (2015). Rating Scales in Survey Research: Using the Rasch model to illustrate the middle category measurement flaw. *Survey Practice 8*(1). https://doi.org/10.29115/SP-2015-0001

Brison, C., Zech, E., Jacken, M., Priels, J-M., Verhofstadt, L., Van Broeck, N., & Mikolajczak, M. (2015). Encounter groups: do they foster psychology students' psychological development and therapeutic attitudes? *Person-Centered & Experiential Psychotherapies, 14*(1), 83-99. https://doi.org/10.1080/14779757.2014.991937

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Routledge

Elliott, R. (2001). The effectiveness of humanistic therapies: A meta-analysis. In D. Cain, & J. Seeman (Eds.) *Humanistic psychotherapies: Handbook of research and practice* (pp. 57-82). American Psychological Association.

Elliott, R., & Greenberg, L. S. (2021). *Emotion-Focused Counselling in Action*. Sage.

Elliott & Rodgers, B. (2012). *Strathclyde Inventory – 16 item version (Version 6).* Unpublished data analyses, University of Strathclyde.

Elliott. R., Wagner, J., Sales, C. M. D, Rodgers, B., Alves, P., & Café, M. J. (2016). Psychometrics of the Personal Questionnaire: A client-generated outcome measure. *Psychological Assessment, 28*(3), 263-278. https://doi.org/10.1037/pas0000174

Folkes-Skinner, J. A. (2011). *A mixed method study of how trainee counsellors change.* Unpublished PhD Dissertation, University of Leicester.

Freire, E. S. (2007). *The Strathclyde Inventory: A psychotherapy outcome measure based on the person-centred theory of change*. Unpublished MSc dissertation, University of Strathclyde.

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting & Clinical Psychology, 59*(1), 12-19. https:// doi.org/10.1037/0022-006X.59.1.12

Johnson, D. A., Knight, D. N., & McHugh, K. (2021). Score reliability and validity evidence for the State-Interpersonal Reactivity Index: A multidimensional assessment of in-session counselor empathy. *Measurement and Evaluation in Counseling and Development, 54*(1), 24-41. https://doi.org/10.1080/07481756.2020.1745652

Lambert, M. J. (2013). The efficacy and effectiveness of psychotherapy. In M. J. Lambert (Ed.). *Bergin & Garfield's handbook of psychotherapy & behavior change* (pp.169-218). John Wiley & Sons, Inc.

Levitt, H. M., Stanley, C. M., Frankel, Z., & Raina, K. (2005). An evaluation of outcome measures used in humanistic psychotherapy research: Using thermometers to weigh oranges. *The Humanistic Psychologist, 33*(2), 113-130. http://doi.org/10.1207/s15473333thp3302_3

Linacre, J. M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement, 3*(2), 103-122.

Linacre, J. M. (2017). A user's guide to *WINSTEPS Ministep Rasch-model computer programs: Program manual 4.0.0).* http://www.winsteps.com

Mearns, D., Thorne, B., & McLeod., J. (2013). *Person-centred counselling in action*. Sage.

McCrae, R. R., & Costa, P. T., Jr. (2013). Introduction to the empirical and theoretical status of the five-factor model of personality traits. In T. A. Widiger & P. T. Costa, Jr. (Eds.), *Personality disorders and the five-factor model of personality* (pp. 15-27). American Psychological Association. https://doi.org/10.1037.13939-002

Murphy, D., Joseph, S., Demetriou, E., & Karimi-Mofrad, P. (2020). Unconditional positive regard, intrinsic aspirations, and authenticity: Pathways to psychological wellbeing. *Journal of Humanistic Psychology, 60*(2), 258-279. https://doi.org/10.1177/0022167816688314

Murray, A. L., McKenzie, K., Murray, K., & Richelieu, M. (2018). Examining response shifts in the Clinical Outcomes in Routine Evaluation- Outcome Measure. *British Journal of Guidance & Counselling, 48*(2), 276-288. https://doi.org/10.1080/03069885.2018.1483007

Owen, J., Adelson, J., Budge, S., Wampold, B., Kopta, M., Minami, T., & Miller, S. (2015). Trajectories of change in psychotherapy. *Journal of Clinical Psychology, 71*(9), 817-827. https://doi.org/10.1002/jclp.22191

Rogers, C. R. (1951). *Client-centred counseling*. Constable.

Rogers, C. R. (1959). A theory of counseling, personality, and interpersonal relationships, as developed in the client-centered framework. In S. Koch (Ed.), *Psychology: A study of a science, volume 3, Formulations of the person and the social context* (pp.184-256). McGraw-Hill.

Rogers, C. R. (1961/1967). *On becoming a person: A therapist's view of psychotherapy.* Constable.

Rogers, C. R. (1963). The concept of the fully functioning person. *Psychotherapy: Theory, Research & Practice, 1*(1)*,* 17-26. https://doi.org/10.1037/h0088567

Rolstad, S., Adler, J., & Rydén, A. (2011). Response burden and questionnaire length: Is shorter better? A review and meta-analysis. *Value in Health, 14*(8), 1101-1108. https://doi.org/10.1016/j.jval.2011.06.003

Saks, J., Fuller, A., Erford, B. T., & Bardhoshi, G. (2020). Meta-study of *Measurement and Evaluation in Counseling and Developme*nt (MECD) Publication patterns from 2000–2019. *Measurement and Evaluation in Counseling and Development, 53*(4), 279-288. https://doi.org/10.1080/07481756.2020.1735222

Skre, I., Friborg, O., Elgarøy, S., Evans, C., Myklebust, L. H., Lillevoll, K., Sørgaard, K., & Hansen, V. (2013). The factor structure and psychometric properties of the Clinical Outcomes in Routine Evaluation –Outcome Measure (CORE-OM) in Norwegian clinical and non-clinical samples. *BMC Psychiatry, 13*(99). https://doi.org/10.1186/1471-244X-13-99

Smith, A. B., Rush, R., Fallowfield, L. J., Velikova, G., & Sharpe, M. (2008). Rasch fit statistics and sample size considerations for polytomous data. *BMC Medical Research Methodology, 8*(33). https://doi.org/10.1186/1471-2288-8-33

Solís Salazar, M. (2015). The dilemma of combining positive and negative items in scales. *Psicothema, 27*(2), 192-199. https://doi.org/10.7334/psicothema2014.266

Spurgeon, S. L., (2017). Evaluating the unintended consequences of assessment practices: Construct irrelevance and construct underrepresentation. *Measurement and Evaluation in Counseling and Development, 50*(4), 275-281. https://doi.org/10.1080/07481756.2017.1339563

Stevens, F. L. (2017). Authenticity: A mediator in the relationship between attachment style and affective functioning. *Counselling Psychology Quarterly, 30*(4), 392-414. https://doi.org/10.1080/09515070.2016.1176010

Warner, M. S. (2017). A person-centred view of human nature, wellness and psychopathology. In S. Joseph (Ed.). *The handbook of person-centred counseling and mental health: Theory, research and practice* (pp. 92-115). PCCS Books Ltd.

Zech, E., Brison, C., Elliott, R., Rodgers, B., & Cornelius-White, J. H. D. (2018). Measuring Rogers' conception of personality development: Validation of the Strathclyde Inventory-French version. *Person-Centered & Experiential Psychotherapies, 17*(2), 160-184. https://doi.org/10.1080/14779757.2018.1473788

Table 1: SI-16 and SI-12 items.

Items	SI-16	SI-12
I have been able to be spontaneous	1	1
I have condemned myself for my attitudes or behavior (R)	2	2
I have tried to be what others think I should be (R)	3	3
I have trusted my own reactions to situations	4	4
I have experienced very satisfying personal relationships	5	
I have felt afraid of my emotional reactions (R)	6	
I have looked to others for approval or disapproval (R)	7	
I have expressed myself in my own unique way	8	
I have found myself "on guard" when relating with others (R)	9	5
I have made choices based on my own internal sense of what is right	10	
I have listened sensitively to myself	11	6
I have lived fully in each new moment	12	8
I have hidden some elements of myself behind a "mask" (R)	13	10
I have felt true to myself	14	
I have been able to resolve conflicts within myself	15	
I have felt it is all right to be the kind of person I am	16	12
I have felt myself doing things that were out of my control (R)		7
I have been aware of my feelings		9
I have felt myself doing things that are out of character for me (R)		11

Note. Items marked (R) indicate those negatively worded items that require reverse scoring.

Table 2. Study 1 datasets presented by protocol, SI version and data collection points.

Dataset	Full (*N* = 1174)				Independent (*N* = 385)			
Protocol	PB	%	SA	%	PB	%	SA	%
Total *N*	846	72.1	328	27.9	294	76.4	91	23.6
SI-31	436	37.1	216	18.4	158	41.0	58	15.1
SI-16	410	34.9	112	9.5	136	35.3	33	8.6
Pre-therapy	302[a]	35.7	96 [a]	29.3	168	57.1	36	39.6
1st session	65	7.7	46	14.0	13	4.4	8	8.8
Mid-1	132	15.6	66	20.1	33	11.2	18	19.8
Mid-2	91	10.8	0	0	19	6.5	0	0
Mid-3	61	7.2	5	1.5	10	3.4	0	0
Mid-4	15	1.8	3	0.9	3	1.0	0	0
Mid-5	5	0.6	2	0.6	0	0	0	0
Mid-6	1	0.1	0	0	0	0	0	0
Post-therapy	106	12.5	61	18.6	28	9.5	17	18.7
Follow up (6 months)	46	5.4	32	9.8	10	3.4	9	9.9
Follow up (18 months)	22	2.6	17	5.2	10	3.4	3	3.3

Notes. PB = practice-based protocol; SA = social anxiety protocol. SI-31 = Strathclyde Inventory (31 item version); SI-16 = Strathclyde Inventory (16 item version); [a] total includes second observations completed by some clients before therapy commenced.

Table 3. Summary of the SI-16 rating scale category structure and fit statistics for persons and items.

Category Label	Score	Count	Mean measure	Infit MNSQ	Outfit MNSQ	Threshold Calibration
Never	0	709	-1.43	.94	.95	None
Only occasionally	1	1662	-.68	.92	.91	-1.86
Sometimes	2	1878	-.06	.93	.93	-.51
Often	3	1286	.63	.94	.95	.61
All or most of the time	4	594	1.33	1.28	1.31	1.76
Persons (N = 385)	30.3	16.0	-.11	1.01	1.01	
SD	11.7	.3	1.07	.60	.58	
Items (N = 16)	728.3	383.1	.00	1.00	1.01	
SD	113.3	.8	.42	.26	.26	

Notes. MNSQ = mean square; *SD* = standard deviation.

Table 4. Correlations and comparison of statistical significance, effect size and correlation of pre-post change on the SI, CORE-OM (psychological distress and risk sub-scales) and PQ.

	Pre-therapy		Post-therapy		Mean difference	*t*	Effect size *d*	CI95(*t*)		Pre/post *r*	Correlation matrix		
	N	*M (SD)*	*N*	*M (SD)*				L	U		SI	CORE-Distress	CORE-Risk
SI	205	1.80 (.66)	205	2.48 (.79)	.68	13.29**	.93	.73	1.14	.50***	-	-	-
CORE-Distress	205	2.07 (.73)	205	1.33 (.82)	.74	13.49**	.95	.75	1.16	.49***	.75**	-	-
CORE-Risk	205	0.38 (.52)	205	0.20 (.43)	.18	5.40**	.38	.18	.57	.51***	.25**	.47**	-
PQ	205	5.13 (.81)	205	3.42 (1.33)	1.72	17.80**	1.55	1.33	1.77	.24***	.54**	.66**	.36**

Notes. Cases excluded listwise if CORE or PQ data missing at pre- or post-therapy. CI95 = 95% confidence intervals; L = lower; U = upper. *** = $p < .001$; ** = $p < .01$ (2-tailed).

```
    ++------+------+------+------+------+------+------+------++
1.0 +                                                         +
    |                                                         |
    |00                                                     44|
    |  00                                                444  |
 .8 +    00                                            44     +
    |      00                                        44       |
    |        00                                     4         |
    |          0                                   4          |
 .6 +           0                                44           +
    |            0                              4             |
 .5 +             00    1                      4              +
    |               *111 111     22     3333334               |
 .4 +             11 0      11222  22233     433              +
    |           11    0     2211    332     4   33            |
    |         11       0  22    1  3   22 44      33          |
    |       11          0*       *3      *2         33        |
 .2 +     11           22 0     3 11    4  2          333     +
    |  111           22    00 33    1144    22           33   |
    |11           222       3*0     4411      222          333|
    |        22222      3333   0***4    1111     22222        |
 .0 +*******************4444444    000000000******************+
    ++------+------+------+------+------+------+------+------++
       -4     -3     -2     -1      0      1      2      3      4
```

Figure 1. Category probability curve of the SI-16 rating scale.

Notes. x-axis = person measure minus item measure (logits); y-axis = probability of response (percentage).

```
-5   -4    -3    -2    -1     0     1     2     3     4     5
|-----+-----+-----+-----+-----+-----+-----+-----+-----+-----| ITEM
0               0    :     1   :   2   :   3    :    4   f { 12-moment
|                                                           |
|                                                           |
0             0    :     1    :  2   :   3     :   4    e  { 9-[NOT]guard
|                                                           |
0            0    :     1   :   2   :   3    :    4        ┌ 13-[NOT}mask
0            0    :     1   :   2   :   3    :    4   d/e  ┤ 15-conflicts
0            0    :     1   :   2  :    3    :    4        └ 7-[NOT]looked
0           0    :     1   :   2   :   3     :   4         ┌ 1-spontaneous
0           0    :     1   :   2   :   3    :    4    c/d  ┤ 2-[NOT]condemn
0           0    :     1   :   2   :   3    :    4         └ 5-relationship
|                                                           |
0          0    :     1   :   2   :   3    :    4       c  { 11-listened
|                                                           |
0         0    :     1   :   2   :   3    :    4           ┌ 14-true
0         0    :     1   :   2   :   3    :    4      b/c  ┤ 16-all right
0         0    :     1   :   2   :   3    :    4           └ 3-[NOT]others
|                                                           |
0        0    :     1    :  2   :   3     :   4            ┌ 4-trusted
0        0    :     1   :   2   :   3     :   4       b/c  ┤ 8-expressed
0        0    :     1   :   2   :   3     :   4            └ 6-[NOT]afraid
|                                                           |
|                                                           |
0    0    :     1   :   2  :    3    :    4             a  { 10-choices
|-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|  ITEM
-5   -4    -3    -2    -1     0     1     2     3     4     5

                       22212322122111
         1 21 1142143944839324890593149528421151          1 1  PERSONS
                T      S     M      S      T
```

Low functioning **Average** **High functioning**

Figure 2. SI-16: Construct KeyMap (expected score item-person matrix).

Notes. “0,1,2,3,4” represents the mean expected score category selected by person according to measure on x-axis; ":" indicates Rasch-half-point threshold; numbers below x-axis are total number of persons (presented vertically) at each measure point; M= mean person measure; S = one standard deviation from mean; T = two standard deviations from mean. Bracketed item or cluster of items indicates 8 strata corresponding to six proposed layers of development: a = self-awareness, b = self-trust, c = self-acceptance, d = openness to self, e = openness to others, f = fully functioning.

```
-5    -4    -3    -2    -1     0     1     2     3     4     5
|-----+-----+-----+-----+-----+-----+-----+-----+-----+-----| ITEM
0                0    :     1    :  2   :   3     :    4 f   8-moment
|                                                           |
0               0    :     1   :   2  :    3    :    4       5-[NOT]guard
|                                                        e  |
0             0    :     1   :   2   :   3     :    4        10-[NOT]mask
0             0    :     1   :   2  :    3    :    4     d   2-[NOT]condemn
0            0    :     1   :   2   :   3     :   4          1-spontaneous
0           0    :     1   :   2   :   3     :    4      c   6-listen
0           0    :     1   :   2  :    3    :    4           12-all right
0          0    :     1   :   2   :   3     :   4        b   3-[NOT]others
0          0    :    1    :  2   :   3     :    4            4-trusted
|                                                           |
|                                                           |
0     0    :     1   :   2   :   3     :    4                7-[NOT]control
|                                                        a  |
0    0    :     1   :   2  :    3    :    4                  9-aware
0   0    :     1   :   2  :    3    :    4                   11-[NOT]character
|-----+-----+-----+-----+-----+-----+-----+-----+-----+-----| ITEM
-5    -4    -3    -2    -1     0     1     2     3     4     5

                         1121111 1 1   1
1           1 1  1214 5389131824819384202142 12 2 1         1  PERSONS
                T      S      M      S      T
```

Figure 3. SI-12: Construct KeyMap (expected score item-person matrix).

Notes. As for Figure 2. Dot/dash lines highlight the boundaries of two standard deviations above and below the mean.

Client ID ____________ Date ___/___/___ Session___________

Please read each statement below and think how often you sense it has been true for you DURING THE **LAST MONTH**. Then mark the box that is closest to this. There are no right or wrong answers – it is only important what is true for you individually.

OVER THE LAST MONTH	Never	Only Occasionally	Sometimes	Often	All or most of the time
1. I have been able to be spontaneous	□0	□1	□2	□3	□4
2. I have condemned myself for my attitudes or behaviour	□4	□3	□2	□1	□0
3. I have tried to be what others think I should be	□4	□3	□2	□1	□0
4. I have trusted my own reactions to situations	□0	□1	□2	□3	□4
5. I have found myself “on guard” when relating with others	□4	□3	□2	□1	□0
6. I have listened sensitively to myself	□0	□1	□2	□3	□4
7. I have felt myself doing things that were out of my control	□4	□3	□2	□1	□0
8. I have lived fully in each new moment	□0	□1	□2	□3	□4
9. I have been aware of my feelings	□0	□1	□2	□3	□4
10. I have hidden some elements of myself behind a “mask”	□4	□3	□2	□1	□0
11. I have felt myself doing things that are out of character for me	□4	□3	□2	□1	□0
12. I have felt it is all right to be the kind of person I am	□0	□1	□2	□3	□4

Thank you for your time in completing this questionnaire

The Strathclyde Inventory: Development of a Brief Instrument for Assessing Outcome in Counseling according to Rogers' Concept of the Fully Functioning Person

Online Supplemental Material

Table of Contents

Table S1. SI-16 Item statistics: misfit order. 1

Table S2. Comparison of SI-16 and SI-20 fit statistics for persons & items. 2

Table S3. SI-20 Item statistics: misfit order. 3

Table S4. Item sensitivity to change. 5

Note S1: Calculation of standardised clinical significance cut-off score and reliable change indices. .. 6

Table S5. Calculation of weighted means. 7

Table S6. Reliable change (p < .05), status change and clinically significant change 8

Table S7. Overview of mid-range SI-20 items to inform removal process. 9

Table S8. SI-20 inter-item correlations (< .6). 10

Note S2: Outline of process and decisions made when reviewing items for removal. 11

Table S9. Comparison of fit statistics for SI-20 and three alternative brief versions: SI-14, SI-13 and SI-12 13

Table S10. SI-12 Item statistics: misfit order. 14

Table S1. SI-16 Item statistics: misfit order.

Item	Measure	Model S.E.	Infit	Outfit		Point Measure Correlation
			MNSQ	MNSQ	ZSTD	
7	.26	.06	1.42	1.51	6.4	.54
6	-.34	.06	1.49	1.45	5.8	.55
5	.12	.06	1.43	1.40	5.2	.58
9	.52	.06	1.09	1.10	1.4	.63
1	.17	.06	1.05	1.09	1.2	.58
3	-.20	.06	1.07	1.07	1.1	.66
10	-1.07	.06	1.06	1.06	.8	.57
13	.32	.06	1.03	1.02	.3	.67
2	.13	.06	.96	.97	-.3	.65
11	-.05	.06	.87	.94	-.9	.68
16	-.18	.06	.85	.85	-2.2	.73
8	-.34	.06	.83	.83	-2.6	.68
12	.80	.06	.81	.79	-3.1	.72
4	-.31	.06	.77	.76	-3.7	.69
14	-.18	.06	.64	.63	-6.1	.77
15	.32	.06	.59	.61	-6.6	.77

Notes. S.E. = Standard Error; MNSQ = Mean Square; ZSTD = *T* statistic;

Table S2. Comparison of SI-16 and SI-20 fit statistics for persons & items.

					Infit		Outfit	
	Score	Count	Measure	Error	MNSQ	ZSTD	MNSQ	ZSTD
SI-16								
Persons (N=385)								
Mean	30.3	16.0	-.11	.34	1.01	-.2	1.01	-.2
SD	11.7	.3	1.07	.08	.60	1.6	.58	1.6
Real RMSE	.35	Adj. *SD*	1.01	Separation	2.92	Person Reliability		.90
Items (N=16)								
Mean	728.3	383.1	.00	.06	1.00	-.3	1.01	-.2
SD	113.3	.8	.42	.00	.26	3.8	.26	3.7
Real RMSE	.06	Adj. *SD*	.42	Separation	6.48	Item Reliability		.98
SI-20								
Persons (N=216)								
Mean	40.2	19.9	-.04	.29	1.01	-.2	1.00	-.2
SD	14.2	.6	1.01	.07	.56	1.6	.52	1.6
Real RMSE	.30	Adj. *SD*	.97	Separation	3.18	Person Reliability		.91
Items (N=20)								
Mean	431.9	214.0	.00	.09	1.00	-.3	1.00	-.2
SD	87.2	1.0	.58	.01	.27	3.1	.27	3.0
Real RMSE	.09	Adj. *SD*	.57	Separation	6.62	Person Reliability		.98

Notes. MNSQ = mean square; ZSTD = standardized mean square; *SD* = standard deviation; Real RMSE = real root mean square error; Adj. *SD* = adjusted standard deviation.

Table S3. SI-20 Item statistics: misfit order.

Item	Measure	Model	Infit		Outfit		PMC
		S.E.	MNSQ	ZSTD	MNSQ	ZSTD	
7	.48	.08	1.39	3.9	1.49	4.7	.52
12	**-.83**	**.08**	**1.45**	**4.3**	**1.41**	**3.8**	**.52**
5	.35	.08	1.38	3.8	1.36	3.6	.59
6	-.15	.08	1.37	3.7	1.33	3.3	.57
18	**-1.20**	**.09**	**1.29**	**2.8**	**1.21**	**1.9**	**.53**
10	-.90	.08	1.14	1.5	1.20	2.0	.54
14	**-1.03**	**.08**	**1.14**	**1.5**	**1.16**	**1.6**	**.42**
15(13)	.46	.08	1.13	1.4	1.12	1.3	.63
1	.29	.08	1.08	.9	1.12	1.3	.55
9	.75	.08	1.02	.3	1.01	.1	.64
3	-.04	.08	.97	-.3	.96	-.4	.68
2	.41	.08	.90	-1.1	.93	-.8	.68
8	-.14	.08	.84	-1.8	.83	-1.9	.67
11	.15	.08	.81	-2.2	.84	-.1.8	.68
20(16)	.09	.08	.79	-2.4	.80	-2.3	.75
19	**.01**	**.08**	**.79**	**-2.5**	**.78**	**-2.5**	**.70**
13(12)	1.01	.08	.73	-3.2	.73	-3.1	.72
4	-.10	.08	.67	-4.1	.66	-4.2	.71
17(15)	.42	.08	.59	-5.3	.60	-5.0	.76
16(14)	-.01	.08	.54	-6.0	.55	-5.9	.78

Notes. S.E. = Standard Error; MNSQ = Mean Square; ZSTD = *T* statistic;

```
-5    -4    -3    -2    -1     0     1     2     3     4     5
|-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|  NUM   ITEM
0                 0    :    1   :   2   :   3    :    4 f  { 13-lived fully in each new moment
|                                                           |
|                                                           |
0               0    :     1   :  2   :   3     :    4  e  { 9-[NOT] found myself "on guard" relating with others
|                                                           |
|                                                           |
0              0    :    1   :   2   :   3    :    4         7-[NOT] looked to others for approval or disapproval
0              0   :     1   :   2  :   3     :    4         15-[NOT] hidden elements of myself behind a "mask"
0             0    :     1   :  2   :   3     :    4   d/e   17-able to resolve conflicts within myself
0             0    :     1   :  2   :   3     :    4         2-[NOT] condemned myself for my attitudes/behavior
0             0    :    1   :   2   :   3     :   4          5-experienced very satisfying personal relationships
0             0   :     1   :   2  :   3     :    4          1-able to be spontaneous
|                                                           |
0            0    :    1   :   2   :   3    :    4           11-listened sensitively to myself
0           0    :     1   :   2  :   3     :    4           20-felt it is all right to be the kind of person I am
0           0    :    1   :   2   :   3    :    4            19-accepted my feelings
0           0    :    1   :   2   :   3    :    4     b/c    16-felt true to myself
0           0   :     1   :   2  :    3    :    4            3-[NOT]tried to be what others think I should be
0          0    :     1   :  2   :   3     :    4            4-trusted in my own reactions to situations
0          0    :    1   :   2   :   3     :   4             8-expressed myself in my own unique way
0          0    :    1   :   2   :   3    :    4             6-[NOT] felt afraid of my emotional reactions
|                                                           |
|                                                           |
0      0    :    1   :   2   :   3    :    4                12-[NOT] felt myself doing things out of my control
0     0    :     1   :   2  :   3     :    4                 10-made choices based on internal sense of right
|                                                       a    |
0     0   :     1   :   2  :    3    :    4                 14-aware of my feelings
|                                                            |
0    0   :     1   :   2  :    3    :    4                  18-[NOT] felt myself doing things out of character
|-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|  NUM   ITEM
-5    -4    -3    -2    -1     0     1     2     3     4     5

                         1111211111
         1  1   1 215215932652732306587216 311121           1  PERSONS
                 T      S     M      S     T
```

Figure S1. SI-20: Construct KeyMap (expected score item-person matrix).

Notes. “0,1,2,3,4” represents the mean expected score category selected by person according to measure on x-axis; ":" indicates Rasch-half-point threshold; numbers below x-axis are total number of persons (presented vertically) at each measure point; M= mean person measure; S = one standard deviation from mean; T = two standard deviations from mean; bracketed item or cluster of items indicates proposed layers of development. **Bold** = items added to create SI-20.

Table S4. Item sensitivity to change.

Item		Pre-therapy		Post-therapy		*t*	*d*	CI95	
	N	*M*	*SD*	*M*	*SD*			L	U
1	225	1.59	1.07	2.16	1.10	7.66**	.53	.34	.71
2	223	1.53	1.09	2.38	1.12	9.76**	.77	.58	.96
3	224	1.76	1.18	2.55	1.16	9.35**	.68	.48	.86
4	224	1.98	1.02	2.60	1.03	8.40**	.60	.41	.79
5	223	1.77	1.19	2.27	1.24	6.22**	.41	.22	.60
6	225	1.83	1.22	2.63	1.20	8.42**	.66	.47	.85
7	224	1.40	1.10	2.24	1.13	10.35**	.75	.56	.94
8	223	2.11	1.12	2.66	1.12	6.98**	.49	.30	.68
9	225	1.24	1.03	2.12	1.17	11.31**	.80	.60	.99
10	225	2.59	1.02	2.84	.97	3.30*	.25	.07	.44
11	224	1.79	1.08	2.47	1.00	8.89**	.65	.46	.84
12[a]	92	2.40	1.26	3.11	1.07	4.61**	.61	.31	.90
13	223	1.11	1.00	1.91	1.12	10.31**	.75	.56	.94
14[a]	92	2.93	.82	3.16	.75	2.25	.29	.00	.58
15	223	1.39	1.13	2.16	1.16	8.68**	.67	.48	.86
16	222	1.84	1.10	2.57	1.10	9.19**	.66	.47	.85
17	222	1.43	.98	2.29	1.05	10.85**	.85	.65	1.04
18[a]	90	2.48	1.06	3.22	.96	5.82**	.73	.43	1.03
19[a]	90	1.87	1.10	2.64	1.12	5.98**	.69	.39	.99
20	221	1.73	1.17	2.58	1.18	9.61**	.72	.53	.91

Notes. [a] = item not included in SI-16. ** = $p < .001$; * = $p < .01$. Cohen's *d*: .20 = small effect; .50 = medium effect; .80 = large effect. CI95 = 95% confidence intervals.

Note S1: Calculation of standardised clinical significance cut-off score and reliable change indices.

Jacobson and Truax (1991, criterion A) proposed that a clinical significance cut-off score can be calculated using data from a dysfunctional population by defining functioning scores as those falling at least two standard deviations in the direction of functionality beyond the mean of the scores collected from the dysfunctional population.

In our study, pre-therapy scores on the Strathclyde Inventory had a mean of 1.79 and a standard deviation of .65 (Table 3). According to Jacobson and Truax criterion A, this indicates a clinical significance cut-off score of 3.09, that is 1.79 + (2 x .65). However, this result is considerably higher than earlier calculations of a clinical significance cut-off score for the Strathclyde Inventory using Jacobson and Truax criteria B and C with data collected in previous studies (Folkes-Skinner, 2011; Freire, 2007; Zech et al., 2018). This is to be expected: Jacobson & Truax (1991) noted that criterion A, calculated using clinical data, tends to produce a more conservative cut-off score, whereas criterion B, calculated using non-clinical data, tends to result in a more lenient result. When both clinical and non-clinical data is available, the preferred method is to use criterion C.

Therefore, we conducted a small meta-analysis of available SI data, by calculating weighted means for SI scores (mean and standard deviation) and, where included, test-retest analyses, for clinical and non-clinical populations (Table S6).

Next, we used these weighted means, first, to calculate a clinical significance cut-off score according to criterion C [(2.79 + 1.94)/2 = 2.36] then, using pooled standard deviations for the whole sample, to calculate RCI minimum value metrics according to the equation:

$$\mathrm{RCI}_{\min} = z\left(s\sqrt{2(1-r_{xx})}\right) *$$

This calculation resulted in minimum RCI values: .96 (*p* < .05) and .64 (*p* < .2).

* where *s* is the weighted mean standard deviation for the non-clinical population (.60); r_{xx} is the weighted mean test-retest for the combined population (.66); and z is the level of *p* required (e.g., for *p* < .05, *z* = 1.96).

Table S5. Calculation of weighted means.

Study	SI mean score				Test-retest		
	N	*M*	WM	*SD*	*N*	T-R	WT-R
Non-clinical samples							
Freire (2007)	399	2.79	1113.21	.54	77	.66	50.82
Folkes-Skinner (2011)	18	2.88	51.84	.51	-	-	-
Zech et al. (2018)	104	2.63	273.52	.48	104	.73	75.92
	119	2.91	346.29	.44	119	.63	74.97
	61	2.73	166.53	.48	-	-	-
	36	2.83	101.88	.86	-	-	-
Total	737		2053.27		300		201.71
Weighted *M*			**2.79**				
Pooled *SD*				.57			
Clinical samples							
Zech et al. (2018)	15	2.13	31.95	.48	9	.46	4.14
	57	2.33	132.81	.71	56	.63	35.28
	10	2.71	27.1	.42	10	.01	.10
This study	225	1.79	402.75	.65	44	.81	35.64
Total	307		594.61		119		75.16
Weighted *M*			**1.94**				
Pooled *SD*				.65			
Whole sample							
Total	1044				419		276.87
Pooled *SD*				**.60**			
Weighted M: T-R							**.66**

Notes. WM = weighted mean of mean; WR-T = weighted mean of test-retest score; Pooled *SD* = pooled standard deviation.

Table S6. Reliable change (p < .05), status change and clinically significant change

	Reliable change (improvement)		Reliable change (deterioration)		No reliable change	
	N	%	N	%	N	%
Total	70	31.1	2	0.9	153	68.0
Status change (N = 93; 41.3%)						
-Clinical to non-clinical	59	26.2*	-	-	29	12.9
-Non-clinical to clinical	-	-	2	0.9	3	1.3
No status change (N = 132; 58.7%)						
-Clinical	6	2.7	-	-	85	37.8
-Non-clinical	5	2.2	-	-	36	16.0

Notes. RCI = .97(*p*<.05); clinical cut-off point = 2.36; * = clinically significant change.

Table S7. Overview of mid-range SI-20 items to inform removal process.

Item	Measure	Misfit[1]	PMC	*d*	DIF	CI-TC	SMC
7. I have looked to others for approval or disapproval (R)	**.48**	**U**	**.52**	**L**	**N**	**.47**	**.46**
15. I have hidden some elements of myself behind a “mask” (R)	.46	G	.63	M	N	.57	.48
17. I have been able to resolve conflicts within myself	**.42**	**O**	**.76**	**L**	**N**	**.74**	**.67**
2. I have condemned myself for my attitudes or behaviour (R)	.41	G	.68	L	N	.64	.55
5. I have experienced very satisfying personal relationships	**.35**	**U**	**.59**	**S**	**N**	**.54**	**.38**
1. I have been able to be spontaneous	.29	G	.55	M	N	.50	.45
11. I have listened sensitively to myself	.15	O	.68	M	N	.66	.61
20. I have felt it is all right to be the kind of person I am	.09	O	.75	M	N	.78	.66
19. I have accepted my feelings	**.01**	**O**	**.70**	**M**	**N**	**.67**	**.59**
16. I have felt true to myself	**-.01**	**O**	**.78**	**M**	**N**	**.76**	**.64**
3. I have tried to be what others think I should be (R)	-.04	G	.68	M	N	.64	.58
4. I have trusted my own reactions to situations	-.10	O	.71	M	N	.68	.54
8. I have expressed myself in my own unique way	**-.14**	**G**	**.67**	**M**	**N**	**.65**	**.54**
6. I have felt afraid of my emotional reactions (R)	**-.15**	**U**	**.57**	**M**	**N**	**.50**	**.44**
12. I have felt myself doing things that were out of my control (R)	-.83	U	.52	M	N	.45	.49
10. I have made choices based on my own internal sense of what is right	**-.90**	**G**	**.54**	**S**	**Y**	**.54**	**.43**

Notes. **Bold** = items removed. [1] = misfit according to infit *z*-scores: U = underfit (*z* > 2.0); G = good fit (-2.0 > *z* < 2.0); O = overfit (*z* < -2.0). PMC = point mean correlation; *d* = effect size of pre-post change: L = large effect (> .75); M = medium effect (> .45); S = small effect (< .44). DIF = differential item functioning; Y = evidence that item function changed between pre- and post-therapy. CI-TC = corrected item-total correlation. SMC = squared multiple correlation; Y = SQM > .9.

Table S8. SI-20 inter-item correlations (< .6).

Item 1	Item 2	*r*
2. I have condemned myself for my attitudes or behaviour	3. I have tried to be what others think I should be	.62
4. I have trusted my own reactions to situations	20. I have felt it is all right to be the kind of person I am	.60
11. I have listened sensitively to myself	**16. I have felt true to myself**	.60
11. I have listened sensitively to myself	**17. I have been able to resolve conflicts within myself**	.68
13. I have lived fully in each new moment.	**16. I have felt true to myself**	.61
13. I have lived fully in each new moment.	**17. I have been able to resolve conflicts within myself**	.63
13. I have lived fully in each new moment.	20. I have felt it is all right to be the kind of person I am	.62
16. I have felt true to myself	**17. I have been able to resolve conflicts within myself**	.62
16. I have felt true to myself	20. I have felt it is all right to be the kind of person I am	.63
17. I have been able to resolve conflicts within myself	20. I have felt it is all right to be the kind of person I am	.66
19. I have accepted my feelings	20. I have felt it is all right to be the kind of person I am	.68

Note. **Bold** = items removed following analysis.

FIGURE 4: STRATHCLYDE INVENTORY – 12 ITEMS

Note S2: Outline of process and decisions made when reviewing items for removal.

Our aim was to create a briefer version of the SI that contained well-fitting, sensitive yet stable items that represented a wide range of item difficulty. We started by thinning out items with closely matched measures of item difficulty: items 7 and 15; items 17 and 2; items 19 and 16; and items 8 and 6.

Item 7 and item 15. Item 7 (*I have looked to others for approval or disapproval*) was an underfit to the model, suggesting responses tended to be erratic. Its PMC (.52) suggested it was not a strong member of the item group. However, it had demonstrated a large pre-post effect size, indicating good sensitivity to change. In comparison, item 15 (*I have hidden some elements of myself behind a 'mask'*) was a good fit to the model, with a moderate PMC (.63) and a medium pre-post effect size. Neither item was correlating with any other item (Table S7). On balance, we decided to remove item 7.

Item 17 and item 2. Item 17 (*I have been able to resolve conflicts within myself*) was overfitting and had one of the highest PMCs (.76), reflected in its presence in four inter-item correlations (Table S7). Item 2 (*I have condemned myself for my attitudes or behaviour*) was a good fit to the model, with a moderate PMC (.68). Both items demonstrated a large pre-post effect size. We discarded item 17.

Item 19 and item 16. Item 19 (*I have accepted my feelings*) was the one of four items returned to SI-20 that had not fulfilled our intended purpose. Instead, it was overfitting, with a moderately high PMC (.70) and medium pre-post effect size, suggesting that it had not contributed to the revised instrument. Item 16 (*I have felt true to myself*) was also an overfitting item, with the highest PMC (.78) in the group, and a medium effect size. Both items correlated with other items (Table S7). For these reasons, we decided to remove both items.

Item 8 and item 6. Item 8 (*I have expressed myself in my own unique way*) was a good fit to the model, with a moderate PMC (.67), while item 6 (*I have felt afraid of my emotional reactions*) was an underfitting item with a relatively low PMC (.57). Both items had medium pre-post effect size. We considered the content of each item, noting that the description 'my own unique way' included in item 8 could be perceived as awkward or alien by some participants, while the experience being described by item 6, might be sufficiently captured by item 12 (*I have felt myself doing things that were out of my control*) and item 4 (*I have trusted my own reactions to situations*). While both items had apparent problems, we made the decision to remove item 6 in the first instance, and to reserve item 8 as a candidate for possible removal.

Item 5. Having completed a review of items with closely matching measures, we considered the statistics for individual items within the middle range of the instrument. Item 5 (*I have experienced very satisfying personal relationships*) stood out as the remaining underfitting item. It had a relatively low PMC (.59) and a small pre-post effect size. We noted the content of the item was more general and less indicative of the theoretical construct than other remaining items. Having considered this range of evidence, we removed item 5.

Item 10. Although reluctant to remove items from the lower end of item difficulty, we reviewed the data collected on item 10 (*I have made choices based on my own internal sense of what is right*), having held doubts about its theoretical fit since our original analysis of SI-16. This item still seemed out of place. In addition, it had a relatively low PMC (.54), a small pre-post effect size, and DIF evidence that the functioning or meaning of this item changed, becoming more difficult, for participants at post-therapy. Having taken all of these points into consideration, we decided to reserve item 10 as a candidate for possible removal.

Table S9. Comparison of fit statistics for SI-20 and three alternative brief versions: SI-14, SI-13 and SI-12.

	SI-20	SI-14	SI-13	SI-12
Person reliability	.91	.87	.86	.85
Person separation	3.18	2.64	2.50	2.42
Person strata	3.91	3.19	3.00	2.89
Item reliability	.98	.98	.98	.98
Item separation	6.62	7.68	7.94	7.83
Item strata	8.49	9.91	10.25	10.11
Variance explained by the measure (%)	61.1	64.1	64.1	63.0

Table S10. SI-12 Item statistics: misfit order.

Item	Measure	Model	Infit		Outfit		Point
		SE	MNSQ	ZSTD	MNSQ	ZSTD	Measure Correlation
7(12)	-.85	.08	1.39	3.7	1.35	3.3	.56
11(18)	-1.23	.09	1.26	2.5	1.17	1.7	.56
9(14)	-1.05	.09	1.17	1.8	1.21	2.0	.44
1	.30	.08	1.06	.7	1.09	1.0	.57
10(15)	.48	.08	1.09	.9	1.07	.8	.65
5(9)	.78	.08	.97	-.3	.96	-.4	.66
3	-.03	.08	.92	-.8	.91	-1.0	.70
6(11)	.16	.08	.87	-1.4	.91	-1.0	.66
12(20)	.09	.08	.88	-1.4	.90	-1.1	.72
2	.42	.08	.87	-1.5	.90	-1.1	.70
8(13)	1.05	.09	.78	-2.5	.80	-2.2	.69
4	-.10	.08	.72	-3.4	.72	-3.4	.69

Notes. *SE* = Standard Error; MNSQ = Mean Square; ZSTD = *T* statistic;