Mindfulness (2021) 12:636–645
https://doi.org/10.1007/s12671-020-01522-3

ORIGINAL PAPER

Applying Generalizability Theory to the Self-Compassion Scale to Examine State and Trait Aspects and Generalizability of Assessment Scores

Oleg N. Medvedev[1] · Anastasia T. Dailianis[2] · Yoon-Suk Hwang[3] · Christian U. Krägeloh[2] · Nirbhay N. Singh[4]

Accepted: 12 October 2020 / Published online: 22 October 2020

Abstract

Objectives Self-compassion has become increasingly recognized for its benefits to human health and well-being, but continued work is required to achieve its accurate assessment in the context of research and psychological treatment. The Self-Compassion Scale (SCS) is widely referred to as a trait measure, although the scale's ability to differentiate between state and trait has not been thoroughly examined using suitable methods. The aim of this study was to use generalizability theory (G Theory) to distinguish state and trait aspects and evaluate generalizability of assessment scores.

Methods The present study utilized G Theory in a longitudinal nested design, using an existing dataset of Australian-based educators ($n = 124$) randomized into intervention (mindfulness/self-compassion) and control group and assessed using the SCS-Short Form (SCS-SF) at baseline, after 8 weeks, and after 6 weeks.

Results The robust psychometric properties of the SCS-SF as a trait measure were confirmed by strong temporal stability and generalizability of scores across occasions and sample population ($G = 0.89$) with only a minor proportion of state-related variance (SCI = 0.11). The majority of individual items clearly measured self-compassion as a trait.

Conclusions The study findings unequivocally demonstrated that the SCS-SF is a valid and reliable measure of trait self-compassion with scores generalizable across sample population and occasions. If significant changes of the SCS-SF scores are observed in individuals as a result of intervention, such changes are likely to be long lasting.

Keywords Mindfulness · State and trait · Measurement · Self-Compassion Scale · Generalizability theory

Self-compassion is a recently introduced psychological construct that emerges from, and builds on, the practice of mindfulness-based therapies (Neff and Germer 2018). Essentially, self-compassion entails extending the same kindness and understanding to oneself as one would offer to a good friend (Neff 2003a). Similar to mindfulness, the self-compassion construct has its roots in Buddhist meditation, and has become widely recognized for its benefits to human health and well-being (Neff et al. 2018). Neff (2003a) began researching the mechanisms and benefits of self-compassion as a healthy alternative to self-esteem, which has a number of pitfalls such as a tendency towards self-centeredness and a lack of concern for others.

Self-compassion is defined as the nonjudgmental awareness of one's own suffering and failures, and the recognition of these as an inherent part of being human (Neff 2003a). There are three interlocking components that make up the construct, namely, self-kindness vs. self-judgement, common humanity vs. isolation, and mindfulness vs. over-identification with painful feelings (Neff 2003a). Self-kindness entails understanding and being patient with oneself in times of failure, adversity, and suffering, rather than being harsh and judgmental. Common humanity recognizes that everyone makes mistakes and we are not alone in our suffering which is, in fact, a normal part of being human. This reduces the sense of isolation experienced by most people when they are facing adversity. And the final concept of increased mindfulness denotes a nonjudgmental awareness that

✉ Oleg N. Medvedev
oleg.medvedev@waikato.ac.nz

[1] School of Psychology, University of Waikato, Private Bag 3105, Hamilton 3240, New Zealand

[2] Auckland University of Technology, Auckland, New Zealand

[3] Centre for Disability Studies, University of Sydney, Camperdown, NSW 2050, Australia

[4] Medical College of Georgia, Augusta University, Augusta, GA, USA

allows both positive and negative feelings to occur without evaluation or censorship (Neff 2003a). All of these components coalesce and interact to form an overall self-compassionate mental attitude (Neff et al. 2017). Learning to develop this mental attitude and consistently practicing kindness towards oneself when faced with life's inevitable difficulties is proposed as a highly adaptive psychological stance (Neff 2003b; Neff et al. 2017).

The purpose of self-compassion training is to cultivate an enduring mental habit that becomes enmeshed in a person's cognition, enabling them to consistently cope better with stress and emotionally challenging situations (Neff 2003a). Successful training would therefore aim to produce long-lasting (trait) changes in participants' cognition, emotions, and behavior rather than transient state changes (Medvedev et al. 2017). A state refers to how an individual is feeling or responding right now, and a trait refers to how an individual feels or responds generally or most of the time. The first is reflecting the person's response to a given situation, and does not last. In contrast, the second is enduring, stable patterns of responding that reflects an individual's personality (Hamaker et al. 2007) and is resistant to change (Epstein 1984). It is now generally accepted that there is an interplay between an individual's situational response and their enduring characteristics and that, when measuring a construct such as self-compassion, both state and trait will contribute to the immediate manifestation of the construct (Hamaker et al. 2007).

The 26-item, self-report Self-Compassion Scale (SCS; Neff 2003b) is widely referred to as a trait measure of self-compassion (Breines and Chen 2013; Neff 2003b; Neff et al. 2017), although this claim has not yet been supported using appropriate and robust methodology. The SCS consists of six separate subscales that reflect the core components of the construct, three positive (self-kindness, common humanity, mindfulness) and three negative (self-judgement, isolation, over-identification). These subscales combine to give an overall total self-compassion score, as well as the six separate subscale scores. It is currently the most widely used scale to assess self-compassion, enabling empirical measurement of the construct as outlined by Neff 2003a, b). While the research on the scale has made progress to date on, for example, establishing validity, internal consistency, and confirming the factor structure of the scale (Neff et al. 2019), a detailed analysis has not yet been undertaken of the performance of the SCS along the state-trait continuum. It has recently been argued that generalizability theory (G Theory) is a particularly suitable method to disentangle the state and trait characteristics of a scale through its ability to accurately examine specific sources of error variance (Medvedev et al. 2017; Truong et al. 2020).

Conventional classical test theory (CTT) methodology evaluates state and trait scales using test-retest correlations, with coefficients below 0.60 indicating a state scale and above 0.70 a trait (Ramanaiah et al. 1983; Spielberger et al. 1970; Spielberger 1999). The major limitation of this method is not accounting for variability of individual items over time and interactions between persons, occasions, and items. A more advanced CTT method to establish temporal reliability employs the intraclass correlation coefficient (ICC) that similarly does not control for variability across individual items and interactions between persons, occasions, and items, which compromises its accuracy (Bloch and Norman 2012; Medvedev et al. 2017). While conventional approaches that emerge from CTT, such as structural equation modelling (Hamaker et al. 2007), can separate the true source of variance caused by the variable of interest (in this case, a person's self-compassion score) from error variance caused by extraneous variables, they are less suitable to identify and evaluate specific sources of error affecting accuracy of assessments. Rather, error is considered as a single factor (Bloch and Norman 2012), which limits possibilities to enhance precision and generalizability of assessment scores to wider populations and contexts.

G Theory provides a robust methodology to investigate trait and state aspects of a construct (e.g., self-compassion) and clearly quantify various sources of the measurement error (Medvedev et al. 2017; Shavelson et al. 1989). G Theory estimates unique contributions of particular sources such as persons, items, occasions, and their interactions to the overall variance and assesses their influences on the overall reliability (Cronbach et al. 1963). The estimated variance components can be utilized to evaluate the contribution of a person representing a trait, and a person-occasion interaction reflecting a state to the overall assessment scores (Medvedev et al. 2017). The ratios of state variance to the summed state and trait variance can be used to reliably differentiate between trait and state aspects in a measure (Medvedev et al. 2017; Paterson et al. 2018).

Given the importance of developing long-lasting habits of self-compassion to deliver genuine positive outcomes for clients, it will be beneficial to understand more clearly how the overall scale, subscales, and even items of the SCS are working in terms of their state and trait characteristics. This may lead to more accurate assessment of interventions and precision of self-compassion measurement in the future. G Theory provides a profile of coefficients that permits differentiation between the stable and dynamic aspects of a measure and accounts for numerous sources of measurement error, namely those related to person factors, item factors, occasion factors, as well as their interactions (Medvedev et al. 2017; Truong et al. 2020).

The current study used G Theory to examine state and trait components of self-compassion and evaluate the temporal reliability and generalizability of the SCS-SF scores. The study included two parts, a generalizability (G) study and decision (D) study. The G study investigated the overall generalizability of the SCS-SF scores and estimated a generalizability (G) coefficient, which is the ratio of the true (person) variance to

the overall variance of scores (Cardinet et al. 2010). The subsequent D study examined characteristics of individual items (e.g., state vs. trait) and item combinations to optimize assessment reliability (Medvedev et al. 2017; Shavelson et al. 1989).

Method

Participants

The participants consisted of 124 educators including $n = 55$ in the intervention and $n = 69$ in the control group who completed all 3 assessments. In order to run G Theory analysis, an equal number of participants in each group are required with all participants providing responses at all 3 time points. After excluding participants with missing data at one of the time points, we randomly removed some participants from the dataset leaving equal groups of 47, nested into two groups, a control group ($n = 47$) and intervention group ($n = 47$), who all completed the SCS-SF at three separate time points. The majority of the participants were female ($n = 85$) with only nine males. However, the male participants were evenly spread across the control group ($n = 4$) and the intervention group ($n = 5$), with the same gender ratio in each group ($\chi^2(1) = 0.12$, $p < 0.05$). The mean age was 45.39 (SD = 12 years) with the majority of the participants between 33 and 57 years of age.

Procedures

Participants were randomly assigned to either the intervention or wait-list control group. The intervention group participated in an 8-week mindfulness course that was designed to support self-management of stress and increase mindfulness, self-awareness, and emotional regulation (Hwang et al. 2019). Participants completed a number of different measurement scales at pre-intervention (T1), post-intervention (T2), and 6-week post-intervention (T3). The control group continued with teaching as usual and did not receive any interventions during the assessment period (Hwang et al. 2019). The data gathered using the SCS-SF were extracted for G Theory analysis in the present study.

Measures

The SCS-SF is a 12-item version of the full-length 26-item SCS. Six subscales measure different aspects of self-compassion: self-kindness (e.g., "I try to be understanding and patient towards those aspects of my personality I don't like"), self-judgement (e.g., "I'm disapproving and judgemental about my own flaws and inadequacies"), common humanity (e.g., "I try to see my failings as part of the human condition"), isolation (e.g., "When I'm feeling down, I tend to feel like most other people are probably happier than I am"), mindfulness (e.g., "When something upsets me I try to keep my emotions in balance"), and over-identification (e.g., "When I fail at something important to me I become consumed by feelings of inadequacy") (Neff 2003b, p. 231). Participants state to what extent each item is true for them on a 5-point Likert scale ranging from 1 (almost always) to 5 (almost never). The six negatively worded items belonging to the self-judgement, isolation, and over-identified subscales are reverse coded so that higher scores indicate higher levels of self-compassion for all items. The scale yields a total self-compassion score while use of subscale scores was not warranted for the short scale version due to reliability concerns (Raes et al. 2011).

Data Analyses

Descriptive statistics and checks for normality of the data were completed using IBM SPSS 25 prior to running the main G Theory analysis, which was conducted using EduG 6.1-e software (Swiss Society for Research in Education Working Group 2006). This study utilized a repeated measures longitudinal nested design, with person (P) nested in group (G) (intervention or control group), by item (I) by occasion (O), expressed as P:G × I × O. The P, G, and O facets are infinite, and I facet is fixed to the number of items in the SCS-SF. The object of measurement, or facet of differentiation, to use G Theory terminology, in this case was persons, which were not considered as a source of error. The facets of generalization, or sources of variance, were the 12 items of the SCS-SF scale and the three different testing occasions (Bloch and Norman 2012). The SCS-SF data were then put through the G-Study analysis, which distinguished the object of measurement (person) or signal from the other facets potentially producing error or noise (Bloch and Norman 2012).

A standard ANOVA was used to compute the variance for each facet, together with their interactions. A Whimbey's correction coefficient (Cardinet et al. 2010) was applied to the classical ANOVA estimates, in order to estimate variance components, this is expressed as $(N(f) - 1)/N(f)$, where $N(f)$ equals the population size of the f-facet in the G study design which effects finite facets, such as the scale items in this case, but has no effect on facets derived from infinite populations, such as persons.

The object of measurement is separated from the other facets in G-study to calculate the variance components of each facet, together with their interactions. G-coefficients for the object of measurement (person) are generated using the computations developed by Brennan (2001). Both relative and absolute G-coefficients are computed. The relative G-coefficient (Gr) accounts only for variance directly related with the object of measurement (Brennan 2001; Cardinet et al. 2010), whereas the absolute G-coefficient (Ga) or Phi (Φ) considers all other sources of variance (e.g., item ×

occasion interaction) that may indirectly affect the absolute measurement (Cardinet et al. 2010). The absolute G-coefficient is a more conservative measure of reliability (Bloch and Norman 2012) and will be referred to from here onwards. Generally, a higher G-coefficient of > 0.80 indicates that a scale or a specific item is reflecting a trait (Arterberry et al. 2014). If the G-coefficient is lower, it may reflect a state measure; however, this should be checked against the SCI for certainty as the source of measurement error must be identified in order to draw a conclusion about the scale or item. The SCI criterion for a state is SCI > 0.60 (TCI < 0.40), and the criterion to assess a measure as trait is SCI < 0.40 (TCI > 0.60) (Medvedev et al. 2017). A subsequent D-study was conducted to analyze the control group and intervention group separately as well as all the individual items within the scale. Interaction between person and occasion (P × O), which reflects a transient or state component, is utilized to estimate the sensitivity of that specific measure to state changes and produces the resulting SCI based on the criteria already mentioned above. SCI was generated for each individual item on the SCS-SF as well as for the total scale.

Results

The data generally met assumptions of normality with skewness and kurtosis falling within the acceptable range, and the most extreme value being for item 4 at − 1.14. Mean SCS-SF scores and standard deviation for the intervention and control groups are presented in Table 1 together with test-retest and Cronbach's α coefficients across times. Test-retest reliability was lower for the intervention group (0.37–0.74) compared to the control group (0.81–0.85). Internal consistency Cronbach's α was strong for the total sample across three time points. There was significant difference only in the scores of the intervention group between T1 and both T2 and T3 ($p < 0.01$).

Table 1 Mean, standard deviation, test-retest, and Cronbach's α coefficients of the SCS-SF scores for the control ($n = 47$) and intervention ($n = 47$) groups across 3 times

Group	Time	Mean	Std. deviation	α	Test-retest
Intervention	T1	34.23	7.73		–
	T2	38.47*	7.48		0.37
	T3	38.47*	6.98		0.74
Control	T1	36.79	8.94		–
	T2	36.87	7.86		0.81
	T3	36.09	8.62		0.85
Total sample	T1			0.87	–
	T2			0.85	0.60
	T3			0.86	0.81

*$p < 0.01$; test-retest Pearson correlations computed with reference to time 1 (T1)

G-Study

ANOVA estimations for person nested in group (P:G), item (I), occasion (O), and their interactions are shown in Table 2. The column labelled "Corrected" shows the ANOVA results after the application of the Whimbey's correction. These data were then used to compute the variance components in the G-study shown in Table 3, and after controlling for person variance, accurately reflect the unique contribution of each potential source of error variance. It can be seen that that corrected group variance component is zero indicating that group is not contributing any variance to the overall variability of assessment scores. This provides solid justification to conduct the G-study with the full sample data.

Table 3 presents variance components reflecting sources of error after controlling for object of measurement (person) variance together with absolute and relative G-coefficients for the SCS-SF for the full sample. Similar to the basic ANOVA, it shows zero variance attributable to group error. The primary finding is an absolute G-coefficient of 0.89 indicating that the SCS-SF is measuring trait aspects of self-compassion with true individual differences in self-compassion trait explaining 89% of variance in the data. The interaction between person and occasion nested in group (PO:G) shown below reflects the state component of self-compassion and appears to be the only major source of error variance in the SCS-SF data. Other sources of error variance are negligible with only 2.8% arising from occasion (O) error alone and 0.2% from group-occasion (GO) interaction error. The SCI for the overall scale was calculated as 0.11 demonstrating a very low amount of variance reflecting a state component. Together, the high G-coefficient and the low SCI clearly indicate that the SCS-SF is not sensitive to state changes and predominantly measures enduring (trait) patterns of self-compassion.

D-Study

Individual facet analyses were computed for the control (Table 4) and intervention (Table 5) groups. When the control group was analyzed upon removal of the intervention group from the analysis, both absolute and relative G-coefficients remained above 0.80 level (0.84) compared to the total sample (0.89). However, removing the control group resulted in both G-coefficients being higher (0.92 to 0.93), indicating higher temporal stability when assessing the intervention group. SCI was 0.16 for the control group alone and 0.08 for the intervention group. Together, these results indicate that SCS-SF is clearly measuring trait independent from the experimental condition (intervention or control). The only relevant source

Table 2 Analysis of variance (ANOVA) of the SCS-SF with person nested in group by item and by occasion design for the total sample ($n = 94$)

				Components				
Source	SS	df	MS	Random	Mixed	Corrected	%	SE
G	2.45	1	2.45	0.00	0.00	0.00	0.0	0.00
P:G	515.26	92	5.60	0.12	0.14	0.14	10.2	0.02
I	29.50	11	2.68	0.00	0.00	0.00	0.0	0.01
O	4.61	2	2.31	0.00	0.00	0.00	0.1	0.00
GI	12.23	11	1.11	0.00	0.00	0.00	0.0	0.00
GO	1.34	2	0.67	0.00	0.00	0.00	0.0	0.00
PI:G	1525.79	1012	1.51	0.24	0.24	0.24	17.6	0.02
PO:G	111.16	184	0.60	0.00	0.05	0.05	3.7	0.01
IO	285.98	22	13.00	0.12	0.12	0.12	9.1	0.04
GIO	30.91	22	1.41	0.01	0.01	0.01	1.0	0.01
PIO:G	1602.01	2024	0.79	0.79	0.79	0.79	58.3	0.03
Total	4121.23	3383					100%	

G=Group effects; *P:G*=Person nested in Group effect; *I*=Item effect; *O*=Occasion effect; *GI*=Group-Item interaction; *GO*=Group-Occasion interaction; *PI:G*= Person-Item interaction nested in Group; *PO:G*=Person-Occasion interaction nested in Group; *IO*=Item-Occasion interaction; *GIO*=Group-Item-Occasion interaction; *PIO:G*=Person-Item-Occasion interaction nested in Group

of error in each individual group analysis was the interaction between person and occasion reflecting the state component, but its influence on the overall strong reliability was merely negligible.

Table 6 presents the variance components of person, occasion, and their interactions as well as the resultant SCI score for every individual item of the SCS-SF. SCI scores are listed in ascending order for individual items. Eleven out of twelve items in the scale were below 0.60 and are thus measuring more enduring characteristics of self-compassion with only item 10 reaching the SCI threshold for a state measure of 0.60 making it the most sensitive item to dynamic changes. Items 2, 9, and 11 fall close to the 0.50 middle point at 0.55, 0.53, and 0.47, respectively, and items 1 and 12 both score

Table 3 G-study statistics for the total sample ($n = 94$) including variance components for sources of error and G-coefficient for the SCS-SF

Source of variance	Differentiation variance	Source of variance	Relative error variance	% relative	Absolute error variance	% absolute
G	(0.000)		.....		.....	
P:G	0.139		.....		.....	
	.....	I	.....		(0.000)	0.0
	.....	O	.....		0.000	2.8
	.....	GI	(0.000)	0.0	(0.000)	0.0
	.....	GO	0.000	0.2	0.000	0.2
	.....	PI:G	(0.000)	0.0	(0.000)	0.0
	.....	PO:G	0.017	99.8	0.017	97.0
	.....	IO	.....		(0.000)	0.0
	.....	GIO	(0.000)	0.0	(0.000)	0.0
	.....	PIO:G	(0.000)	0.0	(0.000)	0.0
Sum of variances	0.139		0.017	100%	0.017	100%
Standard deviation	0.373		Relative SE: 0.130		Absolute SE: 0.132	
G relative	0.84					
G absolute	0.84					

G=Group effects; *P:G*=Person nested in Group effect; *I*=Item effect; *O*=Occasion effect; *GI*=Group-Item interaction; *GO*=Group-Occasion interaction; *PI:G*= Person-Item interaction nested in Group; *PO:G*=Person-Occasion interaction nested in Group; *IO*=Item-Occasion interaction; *GIO*=Group-Item-Occasion interaction; *PIO:G*=Person-Item-Occasion interaction nested in Group

Table 4 D-study statistics for the control group ($n = 47$) including variance components reflecting sources of error and G-coefficient of the SCS-SF

Source of variance	Differentiation variance	Source of variance	Relative error variance	% relative	Absolute error variance	% absolute
P	0.098		.....		.....	
	.....	I	.....		(0.000)	0.0
	.....	O	.....		0.000	0.5
	.....	PI	(0.000)	0.0	(0.000)	0.0
	.....	PO	0.019	100.0	0.019	99.5
	.....	IO	.....		(0.000)	0.0
	.....	PIO	(0.000)	0.0	(0.000)	0.0
Sum of variances	0.098		0.019	100%	0.019	100%
Standard deviation	0.314		Relative SE: 0.138		Absolute SE: 0.138	
G relative	0.84					
G absolute	0.84					

P=Person effect; *I*=Item effect; *O*=Occasion effect; *PI*= Person-Item interaction; *PO*=Person-Occasion interaction; *IO*=Item-Occasion interaction; *GIO*=Group-Item-Occasion interaction

0.43. These five items are better described as sensitive to both state and trait aspects of self-compassion to a comparable degree. There were no consistent patterns of either state or trait tendencies within any of the subscales (self-kindness, self-judgement, common humanity, etc.) suggesting that the full scale represents the optimal measurement design for assessing trait of self-compassion.

Discussion

The aim of this study was to investigate the state and trait characteristics of the SCS using G Theory, which is the best available statistical method for this purpose. G Theory was also employed to analyze the various sources contributing to measurement error and to establish generalizability of the SCS-SF assessment scores. The results showed that the SCS-SF reliably measures trait characteristics of the self-compassion as reflected by high temporal reliability and generalizability of scores across the sample population and assessment occasions. These results are therefore consistent with the test-retest reliability reported by Neff (2003b) and support the original assertion of the scale as a trait measure. This robust supporting evidence is important, and arguably overdue, as the scale has increasingly become referred to as measuring trait self-compassion in the literature (e.g., Breines and Chen 2013; Robinson et al. 2016; Neff et al. 2018; Svendsen et al. 2016) with no strong quantitative evidence to support this assertion. The detailed analysis provided in this study is therefore a strong basis for using the scale in future interventions that aim to target trait changes in self-compassion. If significant differences in the SCS-SF scores can be observed as a result of intervention, such changes are likely to be long-lasting.

The G-study results showed a high G-coefficient (0.89), which provides the primary evidence that the scale, overall,

Table 5 D-study statistics for the intervention group ($n = 47$) including variance components reflecting sources of error and G-coefficient of the SCS-SF

Source of variance	Differentiation variance	Source of variance	Relative error variance	% relative	Absolute error variance	% absolute
P	0.179		.....		.....	
	.....	I	.....		(0.000)	0.0
	.....	O	.....		0.001	6.1
	.....	PI	(0.000)	0.0	(0.000)	0.0
	.....	PO	0.015	100.0	0.015	93.9
	.....	IO	.....		(0.000)	0.0
	.....	PIO	(0.000)	0.0	(0.000)	0.0
Sum of variances	0.179		0.015	100%	0.015	100%
Standard deviation	0.423		Relative SE: 0.120		Absolute SE: 0.124	
G relative	0.93					
G absolute	0.92					

Table 6 D-study variance components of person, occasion, person-occasion interaction, and the State Component Index (SCI) for all items in the SCS-SF. SCI values of 0.60 of higher indicate a state measure

Item	Description	Person	Occasion	Person × occasion	SCI
5	"I try to see my failings as part of the human condition."	0.49	0.10	0.15	0.23
7	"When something upsets me, I try to keep my emotions in balance."	0.42	0.05	0.23	0.36
8	"When I fail at something that's important to me, I tend to feel alone in my failure."	0.46	0.01	0.26	0.37
6	"When I'm going through a very hard time, I give myself the caring and tenderness I need."	0.47	0.02	0.28	0.38
4	"When I'm feeling down, I tend to feel like most other people are probably happier than I am."	0.43	0.02	0.26	0.38
3	"When something painful happens I try to take a balanced view of the situation."	0.41	0.04	0.26	0.39
12	"I'm intolerant and impatient towards those aspects of my personality I do not like."	0.35	0.01	0.27	0.43
1	"When I fail at something important to me I become consumed by feelings of inadequacy."	0.26	0.07	0.20	0.43
11	"I'm disapproving and judgmental about my own flaws and inadequacies."	0.29	0.05	0.26	0.47
9	"When I'm feeling down I tend to obsess and fixate on everything that's wrong."	0.23	0.08	0.26	0.53
2	"I try to be understanding and patient towards those aspects of my personality I do not like."	0.26	0.00	0.32	0.55
10	"When I feel inadequate in some way, I try to remind myself that feelings of inadequacy are shared by most people."	0.23	0.02	0.35	0.60

(Neff 2003b, p. 231)

measures trait self-compassion, and the results of research utilizing the scale can be generalized to the wider population. This means that the scale can be widely used for reliable assessment of dispositional self-compassion in both intervention studies and normal everyday context. The phase of analysis also clearly identified that the error variance in scores was mainly explained by dynamic, or state, aspects of self-compassion, arising through person-occasion interaction. However, this error variance had negligible effects on the overall strong temporal reliability and generalizability of the SCS-SF scores. Theoretically, a perfectly reliable measure of trait characteristics would show 100% of variance attributable to person variance, or a perfect signal-to-noise ratio of 1 in the CTT language regarding ICC (Arterberry et al. 2014; Bloch and Norman 2012). In this case, 89% of the score is signal and only 11% is noise, so the vast majority of the score is explained by person variance, which is very robust. A perfect trait measure is likely not achievable as some person-occasion, or state-related, error is to be expected due the synergistic interaction between a person's self-compassionate state arising in the moment and their habitual traits of self-responding (Cahn and Polich 2006; Kiken et al. 2015). The G-coefficient result is comparable with the other recent G-studies of trait measures, sitting near the higher end for trait reliability. Paterson et al. (2018) reported an absolute G-coefficient of 0.79 for the Children Depression Inventory (CDI-10) and Arterberry et al. (2014) found G-coefficients ranging from 0.81 to 0.89 for the various Big Five personality scales. However, Medvedev et al. (2018) reported a G-coefficient of 0.98 for the widely used Rivermead Post-Concussion Questionnaire, demonstrating that higher reliability is possible. The low state index SCI for the overall scale (0.11), however, is another clear marker that the overall scale is a trait measure.

Interestingly, this analysis shows no error arising from the scale items themselves, which further serves to validate the scale as a suitable trait measure of self-compassion, and also provides robust evidence supporting the internal consistency of the scale. Error variance due to items can indicate that some items are not measuring the intended construct, or that item difficulty is too high (Arterberry, 2014). In this case, G Theory provides evidence that the SCS-SF items are measuring their target construct, and their meaning is consistently interpreted correctly by the majority of participants.

An interesting finding from the D-study is that when the control group and intervention group were analyzed separately, the intervention group G-coefficient (0.92) was higher than that for the total sample (0.89) reflecting a slight increase in enduring patterns of trait self-compassion for this group, while the G-coefficient for the control group (0.84) dropped slightly. This result could be reflective of a ceiling effect if the SCS-SF was perhaps not sensitive enough to measure improvements in self-compassion at the top end of the scale as participants gradually increased their capacity. However, the data showed

no signs of skewness and met conservative assumptions of normality. Another interpretation could be that, as the intervention group progressed and participants increased their self-compassion generally, they may have transitioned their compassionate self-responding from a transient state to a more enduring trait. This result could be understood in light of research into the development of traits as interdependent with states (Cahn and Polich 2006; Kiken et al. 2015), which indicates that it is through repeated state experiences in training intervention programs that expressions of the target cognitions develop into long-lasting traits. This is in line with the original Buddhist meditation practices, which inform both self-compassion and secular mindfulness training. These meditation practices aim to permanently elevate the practitioner's mind to advanced levels of compassionate altruism until the ultimate release from mental suffering described as nirvana or enlightenment is attained. A Buddhist meditator does not aim to only achieve transient experiences of compassion or loving kindness during the meditative state, but rather to mix their mind with such altruistic objects "like mixing water with water" (Gyatso 1990, p. 381) so that it gradually becomes a mental habit and arises spontaneously. So too, in all psychological practice, clinicians are aiming for clients to develop long-lasting protective behaviors and ways of thinking, which result in long-lasting enhancement of quality of life (Ogden 2012).

The D-study findings also contribute further evidence supporting the properties of the SCS-SF items and indicating that the scale is a valid trait measure of self-compassion. When analyzed using the SCI, none of the items demonstrate strong state measure properties as indicated by an SCI of > 0.60. Item 10 just reached the recommended threshold at 0.60, indicating that this item is quite sensitive to state changes in self-compassion. The remaining items in the SCS-SF all fall below the threshold; however, the degree to which they can be characterized as good measures of trait self-compassion varies. Approximately half of the items clearly measure trait (5, 7, 8, 6, 4, and 3) and the other half shows a degree of state sensitivity (1, 2, 9, 11, and 12). Items that fall close to the mid-point of the scale in a band between 0.40 and 0.60 are generally interpreted as measuring both state and trait but not clearly distinguishing between them. It would not be possible to remove any of the items from the SCS-SF as it is already an abbreviated scale and to do so would compromise the internal validity. However, the lack of any items that strongly measure state, combined with the low SCI for the overall scale (0.11) and the high G-coefficient of 0.89, permits to conclude that overall the SCS-SF is a strong measure of trait self-compassion.

The results provide some elucidation regarding a number of the SCS-SF items. It is worth looking at the four specific items (2, 9, 10, and 11) in detail for a moment, as they may have implications for how to approach self-compassion training going forward. While these items are not strongly measuring state per se, they are obviously more sensitive to state changes. It is possible, therefore, that the aspects of self-compassion they measure are more amendable and may change more quickly during short-term interventions. There was no obvious explanation for this within the subscale structure as all of the ambiguous items are from different subscales: item 11 = self-judgement; item 9 = over-identification; item 2 = self-kindness. It is item 10 ("When I feel inadequate in some way, I try to remind myself that feelings of inadequacy are shared by most people") (Neff 2003b, p. 231), which is the most state-like item in the scale (SCI = 0.60) and belongs to the common humanity subscale. However, item 5 ("I try to see my failings as part of the human condition") (Neff 2003b, p. 231) is the most trait-like item (0.23), and it also belongs to the common humanity subscale. Raes et al. (2011) are clear that the subscales in the SCS-SF do not have the same extent of internal consistency as the full version, due to the manner of item selection which prioritized breadth of content over inter-correlation between subscale items. Our results are consistent with this interpretation and indicate that a thorough analysis at subscale level is not necessary.

One can speculate about the more state-like items in the SCS, revealed by this study, and how each of these aspects could be enhanced through training and exercises. The MSC workbook (Neff and Germer 2018) details many of the exercises from the MSC program for use in a self-help context, and may provide clues to address some of the aspects which were sensitive to dynamic changes. The most state-like item—number 10—for example, which comes from the common humanity subscale, may tap into ideas around self-forgiveness and shame. Specific exercises are outlined in the workbook to address these processes through semi-meditative-guided thought processes, and the aspects measured by item 10 may be more responsive to these. Item 5, on the other hand, which is the most trait-like item, but also from the common humanity subscale, may be more oriented towards a recognition of the equanimity of suffering, for which exercises on giving and receiving compassion may be more appropriate. Speculations such as these require further detailed research, but it is clear that understanding the state and trait nature of specific aspects of a construct, and designing items which are finely tuned to measure these, has wide ranging benefits for the efficacy of interventions.

Utilizing instruments that clearly measure enduring and dynamic aspects of the construct would be also beneficial to monitor students and professionals as they progress in their practice. Much research has already looked at the benefits of compassionate self-responding for counselling and psychology students (Nelson et al. 2018) as well as other health care professionals (Raab 2014). This research has found that in the majority of cases caring professionals have a high propensity to care and attend to the needs of others but fail to turn this caring attitude towards themselves. This can be very detrimental as these professions have deleterious effects which

individuals may not be aware of until they face the extreme of burnout (Nelson et al. 2018).

On a very practical level, the ability of G Theory to simultaneously perform multiple analyses with one study also has the added benefit of increasing the efficiency of research and reducing the burden for participants. For instance, although the data in this study were drawn from research involving an intervention, G Theory could reliably and accurately distinguish state and trait characteristics despite the changes in the data over time. With conventional test-retest reliability, the changes in scores as participants improve over time obscure this information as it is only relying on total scores and cannot drill down into the detailed source of the error. This was the case in the present study, where the test-retest reliability for the intervention group and the total sample reflected poor temporal reliability compared to the control group.

Limitations and Directions for Future Research

It should be noted that the SCS and the SCS-SF measure Neff's (2003a) conception of self-compassion and there are other ways to conceptualize and measure it. Gilbert (1989) proposed the social mentality theory framework, which asserts that the ability to be self-compassionate is an intrapersonal behavior which relies on evolved patterns of care-seeking and caregiving. Conversely, self-criticism emerges from interpersonal social patterns that serve to protect individuals from social threats. The Forms of Self-Criticism and Self Reassurance Scales (Gilbert et al. 2004) were developed to assess these two ways of relating to the self, and it would be interesting to examine these in future research using G Theory to understand their state or trait components and how they specifically differ from the SCS. As mentioned earlier, Gilbert et al. (2017) have developed the Compassion Engagement and Action Scales based on a new model of compassion. This model focuses more on the individual's ability for compassionate engagement and the motivation to alleviate distress in self and others by thinking about and taking action to help.

Strauss et al. (2016) also proposed a framework for the assessment of self-compassion, including five key elements: recognizing suffering; understanding the universality of human suffering; feeling empathy for suffering people; tolerating uncomfortable feelings in response to suffering—remaining open and accepting; and motivation to alleviate suffering. This framework aligns well with Neff's conceptualization, but again the motivation to alleviate suffering is not emphasized as strongly in the SCS and is an area that could potentially be added to this conceptualization in the future.

The present study is limited by the use of data gathered only from the SCS-SF, which does not allow for accurate assessment of the scores at a subscale level. However, due to the high correlation between the SCS-SF and the full version of the scale (Raes et al. 2011), the findings can be potentially representative of the self-compassion construct operationalized by the full SCS version.

Although the control and intervention groups included an even spread of males and females, the majority of the participants were female, which may limit the generalizability of scores across the general population. This might also have implications for the results as females are known to have on average lower levels of self-compassion compared to males (Neff et al. 2017). However, as this study was not assessing efficacy, any potential differences in the results of state-trait assessment of the scale itself would likely be negligible.

Ideally, the study would be conducted using a dataset gathered specifically for this purpose and would include more variation at participant level such as the inclusion of meditation practitioners who could reasonably be assumed to have higher levels of self-compassion, a younger and an older cohort, as well as the inclusion of more males in the sample. A new dataset could also allow for more variation in contextual influences such as times of stress, relaxation, or neutrality as were applied in the study by Medvedev et al. (2017).

Analysis using G Theory at an item level offers interesting pathways for tailored treatment plans. It also provides a robust tool with which to develop specifically state or trait psychometric scales in the future. Although there is some ambiguity in approximately half of the scale items, when taken as a whole, the SCS-SF measures enduring not transient characteristics of self-compassion. These results have given strong support for the test-retest reliability data produced by Neff (2003b) in her original development of the scale. This evidence therefore combines to unequivocally demonstrate the SCS-SF as a valid and reliable measure of trait characteristics of self-compassion.

Acknowledgments We express our gratitude to the Australian educators who participated in this project. Data presented in this paper were collected for a project supported by a medical research grant awarded by Teachers Health Foundation.

Authors' Contributions ONM planned the study, conducted statistical analyses, and wrote the manuscript. ATD planned the study, conducted statistical analyses, and collaborated with writing of the manuscript. Y-SH collaborated with the data collection and planning and edited the manuscript. CUK collaborated with the data analyses, writing, and editing the manuscript. NNS collaborated with the data collection and planning and edited the manuscript. All authors approved the final version of the manuscript for submission.

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflict of interest.

Ethical Approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed Consent Informed consent was obtained from all individual participants.

References

Arterberry, B. J., Martens, M. P., Cadigan, J. M., & Rohrer, D. (2014). Application of generalizability theory to the Big Five Inventory. *Personality and Individual Differences, 69*, 98–103.

Bloch, R., & Norman, G. (2012). Generalizability theory for the perplexed: a practical introduction and guide: AMEE guide no. 68. *Medical Teacher, 34*(11), 960–992. https://doi.org/10.3109/0142159x.2012.703791.

Breines, J. G., & Chen, S. (2013). Activating the inner caregiver: the role of support giving schemas in increasing state self-compassion. *Journal of Experimental Social Psychology, 49*(1), 58–64.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer.

Cahn, B. R., & Polich, J. (2006). Meditation states and traits: EEG, ERP, and neuroimaging studies. *Psychological Bulletin, 132*(2), 180.

Cardinet, J., Pini, G., & Johnson, S. (2010). *Applying generalizability theory using EduG*. London: Routledge Academic.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: a liberalization of reliability theory. *British Journal of Statistical Psychology, 16*(2), 137–163.

Epstein, S. (1984). A procedural note on the measurement of broad dispositions. *Journal of Personality, 52*(4), 318–325.

Gilbert, P. (1989). *Human nature and suffering*. Mahwah: Erlbaum.

Gilbert, P., Clarke, M., Hempel, S., Miles, J. N., & Irons, C. (2004). Criticizing and reassuring oneself: an exploration of forms, styles and reasons in female students. *British Journal of Clinical Psychology, 43*(1), 31–50.

Gilbert, P., & Woodyatt, L. (2017). An evolutionary approach to shame-based self-criticism, self-forgiveness, and compassion. In L. Woodyatt, E. Worthington, M. Wenzel, & B. J. Griffin (Eds.), *Handbook of the psychology of self-forgiveness* (pp. 29–41). Cham: Springer.

Gyatso, G. K. (1990). *Joyful path of good fortune: the complete Buddhist path to enlightenment*. Delhi: Motilal Banarsidass Publishers.

Hamaker, E. L., Nesselroade, J. R., & Molenaar, P. C. (2007). The integrated trait–state model. *Journal of Research in Personality, 41*(2), 295–315.

Hwang, Y. S., Goldstein, H., Medvedev, O. N., Singh, N. N., Noh, J. E., & Hand, K. (2019). Mindfulness-based intervention for educators: effects of a school-based cluster randomized controlled study. *Mindfulness, 10*(7), 1417–1436.

Kiken, L. G., Garland, E. L., Bluth, K., Palsson, O. S., & Gaylord, S. A. (2015). From a state to a trait: trajectories of state mindfulness in meditation during intervention predict changes in trait mindfulness. *Personality and Individual Differences, 81*, 41–46.

Medvedev, O. N., Krägeloh, C. U., Narayanan, A., & Siegert, R. J. (2017). Measuring mindfulness: applying generalizability theory to distinguish between state and trait. *Mindfulness, 8*(4), 1036–1046. https://doi.org/10.1007/s12671-017-0679-0.

Medvedev, O. N., Theadom, A., Barker-Collo, S., & Feigin, V. (2018). Distinguishing between enduring and dynamic concussion symptoms: applying generalisability theory to the Rivermead Post Concussion Symptoms Questionnaire (RPQ). *PeerJ, 6*. https://doi.org/10.7717/peerj.5676.

Neff, K. D. (2003a). Development and validation of a scale to measure self-compassion. *Self and Identity, 2*, 223–250.

Neff, K. D. (2003b). Self-compassion: an alternative conceptualization of a healthy attitude toward oneself. *Self and Identity, 2*, 85–102. https://doi.org/10.1080/15298860309032.

Neff, K., & Germer, C. (2018). *The mindful self-compassion workbook: a proven way to accept yourself, build inner strength, and thrive*. New York: Guilford Publications.

Neff, K. D., Whittaker, T., & Karl, A. (2017). Evaluating the factor structure of the Self Compassion Scale in four distinct populations: is the use of a total self-compassion score justified? *Journal of Personality Assessment, 99*(6), 596–607.

Neff, K. D., Long, P., Knox, M., Davidson, O., Kuchar, A., Costigan, A., Williamson, Z., Rohleder, N., Tóth-Király, I., & Breines, J. (2018). The forest and the trees: examining the association of self-compassion and its positive and negative components with psychological functioning. *Self and Identity, 17*(6), 627–645.

Neff, K. D., Tóth-Király, I., Yarnell, L., Arimitsu, K., Castilho, P., Ghorbani, N., Guo, H. X., Hirsch, J. K., Hupfeld, J., Hutz, C. S., Kotsou, I., Lee, W. K., Montero-Maris, J., Sirois, F. M., de Souza, L. K., Svendsen, J. L., Wilkinson, R. B., & Mantios, M. (2019). Examining the factor structure of the Self-Compassion Scale using exploratory SEM bifactor analysis in 20 diverse samples: support for use of a total score and six subscale scores. *Psychological Assessment, 31*(1), 27–45.

Nelson, J. R., Hall, B. S., Anderson, J. L., Birtles, C., & Hemming, L. (2018). Self-compassion as self-care: a simple and effective tool for counselor educators and counseling students. *Journal of Creativity in Mental Health, 13*(1), 121–133.

Ogden, J. (2012). *Health psychology: a textbook: a textbook*. UK: McGraw-Hill Education.

Paterson, J., Medvedev, O. N., Sumich, A., Tautolo, E. S., Krägeloh, C. U., Sisk, R., McNamara, R. K., Berk, M., Narayanan, A., & Siegert, R. J. (2018). Distinguishing transient versus stable aspects of depression in New Zealand Pacific Island children using generalizability theory. *Journal of Affective Disorders, 227*, 698–704.

Raab, K. (2014). Mindfulness, self-compassion, and empathy among health care professionals: A review of the literature. *Journal of Health Care Chaplaincy, 20*(3), 95–108.

Raes, F., Pommier, E., Neff, K. D., & Van Gucht, D. (2011). Construction and factorial validation of a short form of the self-compassion scale. *Clinical Psychology & Psychotherapy, 18*(3), 250–255.

Ramanaiah, N. V., Franzen, M., & Schill, T. (1983). A psychometric study of the State-Trait Anxiety Inventory. *Journal of Personality Assessment, 47*(5), 531–535.

Robinson, K. J., Mayer, S., Batts Allen, A., Terry, M., Chilton, A., & Leary, M. R. (2016). Resisting self-compassion: why are some people opposed to being kind to themselves? *Self and Identity, 15*(5), 505–524. https://doi.org/10.1080/15298868.2016.1160952.

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist, 44*(6), 922.

Spielberger, C. D. (1999). *Manual for the state-trait anger expression inventory-2*. Odessa: Psychological Assessment Resources.

Spielberger, C. D., Gorsuch, R. L., & Lushene, R. E. (1970). *Test manual for the state trait anxiety inventory*. Palo Alto: Consulting Psychologists Press.

Strauss, C., Taylor, B. L., Gu, J., Kuyken, W., Baer, R., Jones, F., & Cavanagh, K. (2016). What is compassion and how can we measure it? A review of definitions and measures. *Clinical Psychology Review, 47*, 15–27.

Svendsen, J. L., Osnes, B., Binder, P. E., Dundas, I., Visted, E., Nordby, H., et al. (2016). Trait self-compassion reflects emotional flexibility through an association with high vagally mediated heart rate variability. *Mindfulness, 7*(5), 1103–1113.

Swiss Society for Research in Education Working Group. (2006). *EDUG user guide*. Neuchâtel: IRDP.

Truong, Q. C., Krägeloh, C. U., Siegert, R. J., Landon, J., & Medvedev, O. N. (2020). Applying generalizability theory to differentiate between trait and state in the Five Facet Mindfulness Questionnaire (FFMQ). *Mindfulness, 11*, 953–963.