Mindfulness (2017) 8:1036–1046
DOI 10.1007/s12671-017-0679-0

ORIGINAL PAPER

Measuring Mindfulness: Applying Generalizability Theory to Distinguish between State and Trait

Oleg N. Medvedev[1] · **Christian U. Krägeloh**[1] · **Ajit Narayanan**[1] · **Richard J. Siegert**[1]

Published online: 23 January 2017

Abstract Mindfulness can be conceptualized as either a state or a trait, but currently, there is no reliable psychometric method to distinguish clearly between the two in psychological measures. Notwithstanding the clinical effectiveness of mindfulness, any specific element of mindfulness treatment can only be evaluated by comparing state and trait changes using techniques that allow such changes to be measured. Generalizability Theory (GT) is a suitable method to differentiate between state and trait variance components, and its application is illustrated here with an empirical example using the Toronto Mindfulness Scale (TMS). Person × occasion interaction is a marker of individual state changes and should explain the largest amount of variance in a valid state measure. To assess state variability, data were collected on three separate occasions: (i) after a holiday, (ii) immediately after a mindfulness exercise, and (iii) before a stressful event (i.e., exam). Generalizability analysis was applied to examine sources of true and error variances. The TMS captured a larger amount of variance attributed to a state and only a small amount associated with trait mindfulness, which is consistent with the purpose of the measure. This study has demonstrated that Generalizability Theory can be usefully applied to distinguish between state and trait components in a measure, and it is recommended as an appropriate psychometric method to validate state and trait measurement tools. These findings have far-reaching implications to improve the accuracy of the distinction between state and trait in mindfulness measurement and other areas of psychological assessment.

Keywords State mindfulness · Toronto mindfulness scale · Measurement · Generalizability theory · Psychometrics · Validation

Introduction

Mindfulness practice has become popular as a safe, non-invasive method for the management of stress and emotional problems and for the improvement of psychological and physical wellbeing (Baer 2003; Brown and Ryan 2003; Rosenzweig et al. 2010). In the context of psychological treatment, mindfulness can be described as "the non-judgmental observation of the on-going stream of internal and external stimuli as they arise" (Baer 2003, p. 125). Recently, there has been an explosion of interest in the application of mindfulness training to a wide range of psychological and health conditions. There is a rapidly growing evidence base for the clinical application of mindfulness practices for alleviating symptoms and enhancing the coping abilities of people suffering from anxiety, stress, depression, emotional instability, substance abuse, post-traumatic stress disorder (PTSD), borderline personality disorder, psychophysiological disorders, suicidal/self-harm behavior, and chronic pain reduction (Hofmann et al. 2010; Ivanovski and Malhi 2007; Rosenzweig et al. 2010). With the increased application of mindfulness-based interventions, the accurate measurement of both a general tendency to be mindful (a *trait*) and an individual's degree of mindfulness at any particular point in time (a *state*) has become an important clinical and research issue. Also, reliable measurement of state and trait mindfulness is necessary during both therapeutic interventions and

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

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

neurophysiological studies (e.g., EEG) on mindfulness (Cahn and Polich 2006; Chiesa and Serretti 2010).

A trait refers to a relatively stable characteristic or enduring behavioral pattern displayed by a person, while a state represents an individual's experience in a given moment, situation, or condition (Hamaker et al. 2007; Spielberger et al. 1970). Essentially, a state is determined by interaction between person and occasion and reflects an individual's unique adaptation to the present moment and environment (Buss 1989; Epstein 1984). However, reliability and validity of psychological measurements such as mindfulness may be compromised through confounding of mindfulness as a state and a trait. It is important to develop and apply reliable methods for distinguishing between the two, otherwise, therapeutic interventions, for example, cannot be assessed for their effectiveness over time. Mindfulness-based interventions aim at lasting or trait changes, and if only state changes are achieved during treatment, relapse is inevitable. This is because state can be explained as more short-term experience (e.g., immediately after a session), whereas trait refers to a pattern established over the longer term (e.g., lasting beyond completion of a mindfulness program).

Generalizability Theory (GT) is an analytical technique for data acquired using psychometric instruments (e.g., rating scales, performance tests). It is named GT because it estimates the extent to which the influence of any specific source of error variance can be generalized to all possible situations and contexts as opposed to only a limited amount of data obtained from a specific testing situation (Cronbach et al. 1963). GT assesses numerous sources of variance contributing to the measurement error associated with the main variable of interest (e.g., a mindfulness score) (Allal and Cardinet 1976). It represents an extension of classical test theory (CTT), based on the idea that every score consists of both true and error values, but it goes beyond its limited assumption considering error variance as a single factor (Allen and Yen 1979). In naturally occurring environments, there are more factors including personal (e.g., personality), methodological (e.g., psychometric characteristics of the measure used), and situational (e.g., time of the day) that might each independently contribute to measurement error. GT provides an advanced method for assessing these factors and their interactions thus contributing to the improvement of methodology and precision of an assessment instrument.

GT employs repeated measures factorial analysis of variance (ANOVA) to estimate the relative contribution of different sources of variability to the overall measurement error, which is also referred to as "noise" (Brennan 2001). Every such contribution can be expressed as an intra-class correlation coefficient (ICC) ranging from 0 to 1. An ICC is a reliability coefficient that expresses the ratio between the amount of variance in scores attributed to the primary variable being measured and the total amount of observed variance. For instance, the amount of variance between mindfulness scores that is explained by differences between the participants can be represented as an ICC that reflects the discriminative ability of the mindfulness questionnaire as follows (Bloch and Norman 2012):

$$\mathbf{ICC} = \frac{\text{variance (participants)}}{\text{variance (participants)} + \text{variance (error)}}$$

Here, ICC depends on two factors: the actual ability of an instrument to discriminate between participants and the amount of noise due to other influencing factors. ICC was originally introduced in CTT, represented by a slightly different but essentially similar formula using the concept of "signal-to-noise ratio" (SNR) (Fisher 1925). SNR is mathematically equal to the square of the effect size (ES^2), which could be extracted from any ANOVA analysis and represents a ratio between consistent change (variance) in the X variable that refers to ΔX and total variance (σ^2) in the data (Bloch and Norman 2012):

$$\text{SNR} = \text{ES}^2 = \frac{\Delta X^2}{\sigma^2}$$

Therefore, ICC based on SNR definition is expressed by the following formula:

$$\text{ICC} = \frac{\text{SNR}}{1 + \text{SNR}}$$

The larger the amount of variance in a variable of interest (signal) compared to noise, the better are the chances to detect these changes reliably. An ICC closer to 1 would indicate that there is mainly a real difference related to signal and relatively a low amount of noise, and an ICC close to 0 would indicate that there was mainly noise or error in the data. ICC refers to a G-coefficient in GT terminology and similarly expresses the ratio of the observed (true) variance due to the object of measurement (σ^2_{p}) and the total variance of universe scores including the observed (true) variance and the error variance (σ^2_{error}) (Brennan 1992; Shavelson et al. 1989):

$$G_p = \frac{\sigma_p^2}{\sigma_p^2 + \sigma^2_{\text{error}}}$$

A G-coefficient is normally computed for the variable of interest (e.g., trait mindfulness) but can also be computed for every factor contributing to error variance, given that a research design provides relevant data to assess variability due to these contributions (Bloch and Norman 2012). In this case, the G-coefficient expresses the generalizability of influence attributed to specific factors to all possible situations and contexts.

GT can be used to identify and compare the amount of variance uniquely explained by the person, the item, and the

occasion plus their respective interactions (Brennan 2001; Bloch and Norman 2012). The variance due to person-occasion interaction is a direct reflection of the "stateness" of a latent construct, while person variance alone is a representative of a trait (Buss 1989; Chaplin et al. 1988; Epstein 1984). Importantly, GT permits this analysis for the total test, subscales, and even individual items. In other words, true "state items" can be distinguished from items that are not truly sensitive to occasion. Estimation of variance associated with the object of measurement (e.g., persons) and influencing facets (e.g., occasions) is conducted in a G-study (generalizability study). Variance components are estimated based on observed values obtained from the universe of all possible (hypothetical) observations. Scales and individual items measuring state are expected to reflect a higher amount of variance attributed to person-occasion interaction and low generalizability across occasions (e.g., G < 0.70) as opposed to reliable trait measures, which are expected to have G of 0.80 or higher (Arterberry et al. 2014; Gardinet et al. 2009). However, traits are the basic determinants of states through interaction with situational factors for the same latent construct, and a precise distinction between state and trait can only be estimated based on their variance components (Hamaker et al. 2007; Geiser et al. 2015). To date, there are no commonly accepted bench marks for the relative proportions between state and trait components in a valid state measure, and we propose the state component index (SCI) to estimate this relationship as follows:

$$\mathrm{SCI} = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_t^2}$$

In the above formula, the variance component of a state ($\sigma_s^2 = \sigma_{po}^2$) is essentially the noise or error variance due to person-occasion interaction that affects trait scores. This reformulation of the original ratio equation is essentially identifying the ratio of state to trait including noise in both which we can assume be equal because the trait (persons) component ($\sigma_t^2 = \sigma_p^2$) is the basic component of the state variance. To ensure accuracy of measurement, the SCI calculation should use an absolute value of variance due to person-occasion interaction derived from G-analysis that accounts for all sources of error variance identifiable in the data. SCI is developed in line with GT logic and is easy to interpret. For instance, SCI = 1.00 would mean that there is no trait component and only individual state is measured, which appears unlikely because a trait is a basic predictor of a state (Buss 1989; Epstein 1984). SCI = 0.50 would mean that state and trait components are the same, and a scale cannot be classified as either state or trait measure. However, SCI > 0.60 can be considered as a characteristic of a state measure with higher scores corresponding to a better ability of an instrument to capture state changes. Similarly, trait component index (TCI) can be used to validate a trait measure using the same metric:

$$\mathrm{TCI} = \frac{\sigma_t^2}{\sigma_t^2 + \sigma_s^2}$$

Therefore, more precise distinction between scales measuring states and traits can be made based on G-study results. The D-study (decision study) is based on G-study results and involves experimenting with designs (e.g., fixed or random) in an attempt to reduce measurement error (Brennan 2001; Shavelson et al. 1989). It can be used to identify those items that are not consistent with the purpose of the measure (e.g., items measuring trait in a state measure) and thus to improve an instrument by removing them.

The traditional method for demonstrating distinct state and trait components in a scale has been to examine test-retest reliability coefficients, which are expected to be lower for a valid measure of state (e.g., <0.60) and higher for a trait measure (e.g., >0.70) (Ramanaiah et al. 1983; Spielberger et al. 1970, 1999). The main limitation of this method that it is based entirely on the total score correlations at Time 1 and Time 2. If relationships and distinctions between trait and state are to be given a solid, systematic, and robust foundation, there is a need to understand the different contributions made by item effects, scale effects, person effects, and occasion effects to changes in trait and state. Identifying such effects will require a much deeper analysis of variances found in the different dimensions of the research study so that such variances can be identified and isolated if necessary to provide a greater control in future experimental studies. Most importantly, the test-retest coefficient fails to account for variability due to interaction between person and occasion, which is an essential determinant of state changes in an individual (Buss 1989; Chaplin et al. 1988; Epstein 1984). Put simply, we do not expect trait scores to vary a great deal across situations. In contrast, the interaction between the person and the occasion is a state by definition. To date, the exploration of state and trait variability is limited to structural equation modeling (SEM) approaches (e.g., Hamaker et al. 2007; Geiser et al. 2015; Kenny and Zautra 2001) that are generally useful to study state-trait relationships. However, none of the proposed SEM methods account for various sources of variance (e.g., an item) contributing to the measurement error associated with state and trait variability, which limits their applicability for validation of state and trait measures. Such differences in variability require a more detailed study of how factors or components that can affect state and trait, including person and situation, can be quantified so that changes

in state and trait can be predicted by knowing of changes in person and situation, which is a true generalizability, in other words.

While GT was applied to assess reliability of trait measures (e.g., Arterberry et al. 2014), we are not aware of any studies to date that have used GT methods to distinguish between state and trait components in a state measure. The aim of this study is to demonstrate application of GT to investigate state- and trait-related variance components in the Toronto Mindfulness Scale (TMS) (Lau et al. 2006), the first and the most frequently cited instrument designed exclusively to assess state mindfulness. To increase state variability, data were collected on three separate occasions: "after a University holiday", "after a brief mindfulness exercise", and "before a stressful event (class test)". GT analysis was based on the procedure described elsewhere (Bloch and Norman 2012; Gardinet et al. 2009). Two-way repeated measures ANOVA was used in the G-study design to assess the variance due to object of measurement (persons) and sources of error variance due to occasion, item, person-occasion, person-item, and person-occasion-item interactions of the TMS subscales. We expected a low generalizability of individual scores across occasions ($G < 0.70$) (Arterberry et al. 2014) and a high amount of variance due to person-occasion interactions reflected by the proposed SCI above 0.60 as characteristics of a valid state measure. The D-study was conducted to demonstrate how the functioning of the TMS subscales and individual items can be investigated and optimized by varying facets designs.

Method

Participants

The sample size ($n = 55$) satisfied criteria for a reliability study in medical research (Shoukri et al. 2004) and is adequate for generalizability analysis because G-coefficients are essentially similar to reliability coefficients (Bloch and Norman 2012). Given the experimental nature of this study, where the focus is on an initial measurement of the sample followed by an intervention that is subsequently measured, no attempt was made to set up a control group. Also, any biases introduced by the convenience sampling method involved (all participants were locally available and indicated willingness to participate) are assumed to be distributed evenly throughout the sample. All 55 participants, who provided data at three different occasions, were New Zealand university students, (78.2% females, 21.8% males) with a mean age of 23.44 (SD = 6.32) and range of 18 to 44. Ethnic groups include Caucasians (49.1%), Polynesians (16.4%), Asians (14.5%), and other ethnicities (20%).

Procedure

Potential participants were approached in lectures and invited to complete the survey on three different occasions and to hand the survey directly back to the researchers or submit it to a locked collection box at their respective faculty. Three occasions were chosen to increase variability of state mindfulness, and data were collected "after a holiday", "after a mindfulness exercise", and "before a stressful event". On the first occasion, the first lecture after the summer holiday served as the baseline. Here, students completed the questionnaire in class before the lecture or during a short lecture break. The second occasion occurred after a one-week interval, where students completed the questionnaire at the beginning of laboratory classes in a different environment and in smaller groups. Prior to completing the questionnaire on occasion 2, students participated in a 10-min guided mindfulness exercise called "body scan", which is a standard component of Mindfulness-Based Cognitive Therapy (MBCT) (Segal et al. 2013). It was expected that the mindfulness exercise would increase or at least influence mindfulness levels of the participants. To ensure the same conditions across lab classes and to minimize experimenter effects, the "body scan" exercise instructions were played to the participants from the audio CD included in the book *Mindfulness: Finding Peace in a Frantic World* (Williams and Penman 2011). On the third occasion, which occurred after a one-month interval after the first data collection, students completed the questionnaire in the lecture theater before the lecture. This occasion was a week before an important class test, and the lecture included the test overview and relevant discussion. It was expected that students would have higher stress levels on this occasion, which might impact on their mindfulness levels. The students were asked to create a unique ID containing letters and number (e.g., ABC123), which could not be used to identify them but to match the questionnaires completed by the same person on three different occasions. The authors' university ethics committee had approved this study.

Measure

The TMS (Lau et al. 2006) is a 13-item self-report questionnaire designed to measure two dimensions of state mindfulness: curiosity and decentering. The former is defined as present-moment awareness with a quality of curiosity, while the latter refers to awareness of one's experience from a distant observer perspective and thus without identifying oneself with the content of one's thoughts and feelings and getting carried away by them (Lau et al. 2006). Meditators scored higher on both TMS subscales compared to those without meditation experience, and decentering scores were shown to reflect meditation experience (Davis et al. 2009) and changes in psychological symptoms (Lau et al. 2006). Both TMS subscales

displayed increased scores after mindfulness training, which provide support for their construct validity, although no test-retest reliability scores were reported (Park et al. 2013). The TMS includes a 6-item curiosity subscale (Cronbach's alpha 0.86–0.91) and a 7-item decentering subscale (Cronbach's alpha 0.85–0.87) (Park et al. 2013). Both subscales use a 5-point Likert-scale response format (0 = "Not at all" to 4 = "Very much"). The total subscale scores are calculated by adding responses to individual subscale items with higher scores corresponding to higher levels of state mindfulness.

Data Analyses

Descriptive statistics together with Cronbach's alpha coefficients and test-retest bivariate correlations for the curiosity and decentering subscales of the TMS were computed using IBM SPSS version 23 at each of the three assessment occasions. Test-retest reliability scores for a state measure were expected to be in the range from 0.16 to 0.57 (Ramanaiah et al. 1983; Spielberger 1999).

GT analyses were conducted using EduG 6.1-e software (Swiss Society for Research in Education Working Group 2006) that produces an extended output, which is easier to interpret in practical terms. We employed a random effects design with two crossed facets for both G and D-study: persons (P), by occasion (O), by item (I), expressed as P × O × I, where the P and O facets are infinite and the I facet is fixed. The facets were defined from the trait perspective with persons as the object of measurement, which is a facet of differentiation, and items and occasions as instrumentation facets (Gardinet et al. 2009). States are expected to vary across occasions reflected by person-item interaction, but not across items. Here, the error variance attributed to interaction between person and occasion (P × O) will be indicative of a state component in a scale score, which is expected to be relatively strong for a state measure.

Conventional ANOVA was used to compute the sums of squares, mean squares, variance components, and variance percentages associated with each facet including standard errors. Variance components were estimated for each effect based on their mean squares and samples to assess measurement error due to each of the sources using formulas developed by Brennan (1977, 1992). Variance components are estimated by EduG after applying a Whimbey's correction to classical ANOVA estimates that accounts for facets, which are not sampled from infinite universes (e.g., scale items) (Gardinet et al. 2009). It is expressed as ((N(f)−1)/N(f)), where N(f) is the universe size of the f facet in the G-study design and has no effect on merely random facets.

Generalizability analysis was applied to estimate contribution of each facet to variance of universe scores including relative and absolute error variance and to calculate relative and absolute G-coefficients for the object of measurement (persons). Relative G-coefficient only accounts for variance directly influencing a relative measurement tool (e.g., person-occasion and person-item interactions) (Shavelson et al. 1989) and may express commonly used ρ^2, ϖ^2 or intermediate value by the virture of using Wimberley's correction (Gardinet et al. 2009):

$$G_{\text{relative}} = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_\delta^2}$$

Here, σ_p^2 is the variance due to the object of measurement (persons) and $\sigma_\delta^2 = \sigma_{po}^2 + \sigma_{pi}^2 + \sigma_{poi}^2$ is the relative error variance. Absolute G-coefficient ($G_{absolute}$) is similar to the comonly used Phi (Φ) coefficient after applying Wimberley's correction. It accounts for an absolute error variance ($\sigma_\Delta^2 = \sigma_o^2 + \sigma_i^2 + \sigma_{io}^2 + \sigma_{po}^2 + \sigma_{pi}^2 + \sigma_{poi}^2$) that includes other factors (e.g., items and occasions) influencing an absolute measure (Gardinet et al. 2009):

$$G_{\text{absolute}} \approx \Phi = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_\Delta^2}$$

Also, the SCI to estimate relationship between state and trait variance components was computed using the formulae proposed in the introduction. D-study included facets analyses of every individual item to estimate variance components and G-coefficients associated with the object of measurement (persons or trait), and variance due to person-occasion interaction as a state marker. It also involved testing various facet designs by manipulating their levels to optimize the instrument.

Results

All data distributions met normality assumptions with skewness and kurtosis values fairly close to zero and non-significant Shapiro-Wilk normality tests. Repeated measures ANOVA indicated that the effect of occasion was significant for both facets of state mindfulness: curiosity ($F(2, 54) = 6.88$, $p = .002$, $\eta^2 = .11$) and decentering ($F(2, 54) = 12.46$, $p = .001$, $\eta^2 = .19$). Post-hoc tests showed that the mean curiosity and decentering levels on occasion 2 (1 Week, after mindfulness exercise) were significantly higher compared to both other occasions. Table 1 presents descriptive statistics together with Cronbach's alpha coefficients and test-retest bivariate correlations for the curiosity and decentering subscales of the TMS at each of the three assessment occasions. While the curiosity subscale showed good internal consistency at all three occasions, the decentering alpha coefficients varied but in the acceptable range from .70 to .80. According to the expectations for a state measure, test-retest reliability coefficients for both subscales at 1 week and 1 month intervals ranged from .38 to .46 (Table 1).

Table 1 Means, standard deviations (SD), internal and test-retest reliability estimates for the TMS[a] curiosity and decentering subscales (n = 55)

Subscale/measurement	Baseline (in lecture)	1 week (mindfulness exercise)	1 month (before test)
Curiosity			
Mean (SD)	10.04 (5.08)	12.05* (5.73)	8.91 (5.36)
Cronbach's alpha	.83	.87	.88
Test-retest (r)[b]	–	.38	.34
Decentering			
Mean (SD)	10.09 (4.80)	13.44[a] (5.62)	10.36 (5.12)
Cronbach's alpha	.70	.80	.79
Test-retest (r)[b]	–	.44	.46

* Mean is significantly different from two other means ($p < .05$); [a] TMS = The Toronto Mindfulness Scale;

[b] Test-retest bivariate correlations were computed between the baseline scores and scores after 1 week as well as baseline scores with scores after a 1-month interval

G-Study

ANOVA results for the TMS curiosity and decentering subscales together with variance components attributed to person (P), item (I), and occasion (O), and interactions between them are included in Table 1 and provide basic estimates for the G-study. Corrected variance components included in columns 7 and 8 (in %) are computed by applying Whimbey's correction. Relative and absolute contribution of the percentage values presented in column 8 (Table 2) were estimated from a GT perspective and are presented in Table 3. The largest amount of variance of both subscales scores was explained by person-occasion interactions, which is a marker of individual state changes in domains of curiosity and decentering across three different occasions.

The results of a generalizability analysis of both curiosity and decentering TMS subscales are presented in Table 3. Components that cannot be computed (as they did not exist) in the current design are represented as a row of dots. As predicted for a valid state measure, person-occasion (P × O) interaction is the main source of error variance for both subscales explaining over 90% of relative and absolute error variance. The final results are relative, and absolute G-coefficients are both below the acceptable level of 0.80 recommended for the assessment of traits and in line with expectations for a state measure. The proposed SCI values were calculated based on differentiation variance of person (trait: $\sigma_t^2 = \sigma_p^2$) and absolute error variance of person-occasion interaction (state: $\sigma_s^2 = \sigma_{po}^2$) for the curiosity (SCI = 0.70) and for the decentering (SCI = 0.75) subscales. These values indicate that, after accounting for all sources of error identifiable in the data, both subscales mainly reflect variance associated with state changes in line with expectations for a valid state measure.

D-Study

Facets analysis was conducted first, to obtain variance estimates for every individual item by excluding all other items. The estimates for a differentiation facet of a person together with estimates for person-item interaction and G-coefficients are included in Table 4. In line with expectations for items measuring state, most of the items show a high amount of variance attributed to person-item interaction and typically above 0.4 with the exception of items 1 and 11, which are just below this benchmark. Low differentiation estimates (P) were found for most of the items consistently reflected by the low values of the G-coefficients in the right column, which are both expected to be high for a trait measure (i.e., G-coefficient above 0.80).

However, two items, 4 and 7 in the decentering subscale did not reflect any variance attributed to a trait (person) and consequently had generalizability coefficients of zero. Therefore, we tested the relative contribution of these items to the decentering subscale by removing them. After removing those two items, the proportion of variance due to person-occasion interaction decreased from 100% (Table 4) to 79.1% and produced an additional 19.80% error variance attributed to person-occasion-item interaction, which is a threat to scale reliability. Also, removing those items did not affect the G-coefficients remaining at the same level of 0.24 (relative) and 0.23 (absolute). This illustrates that items 4 and 7 contribute to the overall reliability of the decentering subscale in discriminating between state levels. Removing individual items from each subscale did not result in an increase but in some cases decreased the overall generalizability coefficients. Finally, removing Occasion 3 (before the class test) slightly increased G-coefficients in the curiosity subscale up to 0.44 (absolute and relative) and removing occasion 1 (baseline, after the holiday) decreased the overall G-coefficients of both subscales just below 0.1 (absolute and

Table 2 ANOVA for the curiosity (*above*) and decentering (*below*) subscales of the Toronto Mindfulness Scale (TMS) including the sum of squares (SS), degrees of freedom (df), mean squares (MS), variance components (in %), and standard errors (SE) for the Person (P) × Occasion (O) × Item (I) design including interactions (n = 55)

Source	SS	df	MS	Curiosity variance components				
				Random	Mixed	Corrected[a]	%	SE[b]
P	341.44	54	6.32	0.10	0.10	0.10	6.70	0.07
O	1.53	2	0.76	0.00	0.00	0.00	0.00	0.00
I	11.97	5	2.39	0.01	0.01	0.01	0.70	0.01
P × O	489.80	108	4.54	0.64	0.69	0.69	46.50	0.10
P × I	179.20	270	0.66	0.00	0.00	0.00	0.00	0.02
O × I	6.79	10	0.68	0.00	0.00	0.00	0.00	0.01
P × O × I	371.88	540	0.69	0.69	0.69	0.69	46.10	0.04
Total	1402.61	989					100	
				Decentering variance components				
P	267.87	54	4.96	0.06	0.06	0.06	3.70	0.05
O	1.28	2	0.64	0.00	0.00	0.00	0.00	0.00
I	105.89	6	17.65	0.10	0.10	0.09	6.10	0.05
P × O	408.63	108	3.78	0.41	0.54	0.54	31.60	0.07
P × I	280.02	324	0.86	0.00	0.00	0.00	0.00	0.03
O × I	11.34	12	0.94	0.00	0.00	0.00	0.10	0.01
P × O × I	576.76	648	0.89	0.89	0.89	0.89	58.40	0.05
Total	1651.78	1154					100	

[a] Corrected components are calculated by applying Whimbey's correction to the classical ANOVA estimates

[b] SE in the right column is related to the mixed effects presented in the column 6

relative). Removing occasion 2 (1 week, mindfulness exercise) did not result in any substantial changes of the overall G-coefficients.

Discussion

The aim of this study was to demonstrate the application of GT to distinguish between state and trait variance components in a measure using the TMS as an example. This study has demonstrated that GT can be applied to distinguish between state and trait components in a measure, and it is recommended as an appropriate psychometric method to validate state and trait measurement tools. The method and the sequence of analysis illustrated in the "Results" section allows researchers to assess the validity and reliability of any psychometric measure of a state or a trait using GT. Currently, the only statistical method used to distinguish between state and trait measures is merely a correlation between total test scores at two different occasions (test-retest). The proposed GT method is based on an accurate estimation of variance components of both state and trait that accounts for various sources of error variance and provides an advanced method for validation of state and trait measures. It is particularly powerful in its ability to examine the "stateness" or "traitness" of each individual item.

To demonstrate the application of GT, we used a state measure of mindfulness, the TMS (Lau et al. 2006). We chose this measure because, while GT has already been used to assess the reliability of trait measures (Arterberry et al. 2014), it has not previously been used to distinguish between state and trait mindfulness. Before using the TMS to illustrate the application of GT methods, reliability and construct validity of the instrument were tested using more traditional methods and supported by the results (Table 1). Prior to GT analysis, we also ensured that the data met assumptions of normality. Although, not the main purpose of the study, the results provide a support for construct validity of the TMS as a state measure as the scores followed predicted changes, namely, increased mindfulness after a brief mindfulness exercise and decreased mindfulness during a stressful pre-exam period. These findings are consistent with Lau et al. (2006) who reported an increase of the TMS scores following a mindfulness-based intervention.

In this G-study, two-way repeated measures ANOVA was used first to extract the variance due to the object of measurement (persons) reflecting a trait, person-occasion interaction reflecting a state, and other sources of error variance such as occasion, item, and interactions of the TMS subscales. Such ANOVA results are important because they provide basic estimates for further analysis. In terms of a state-trait distinction, a trait measure should have the largest amount of variance explained by the person and a state measure, as in this case, by the person-occasion interaction (Table 2). However, traditional ANOVA is not precise enough to identify such individual contributions. For instance, it can be seen that variances

Table 3 Estimated variance components with standard errors (SE) and G-coefficients and for the G-study P × O × I design including the TMS subscales curiosity (*above*) and decentering (*below*)

Source of variance	Differentiation variance	Relative error variance	% Relative	Absolute error variance	% Absolute
Curiosity TMS subscale[a]					
P	0.10	–		–	
O	–	–		0.00	0.00
I	–	–		0.00	0.40
P × O	–	0.23	91.20	0.23	90.90
P × I	–	0.00	0.00	0.00	0.00
O × I	–	–	–	0.00	0.00
P × O × I	–	0.02	8.80	0.02	8.80
Sum of variances	0.10	0.25	100	0.25	100
Standard deviation	0.32	Relative SE: 0.50		Absolute SE: 0.50	
G relative 0.28					
G absolute 0.28					
Decentering TMS subscale[b]					
P	0.06	–		–	
O	–	–		0.00	0.00
I	–	–		0.00	0.00
P × O	–	0.18	100.00	0.18	100.00
P × I	–	0.00	0.00	0.00	0.00
O × I	–	–	–	0.00	0.00
P × O × I	–	0.00	0.00	0.00	0.00
Sum of variances	0.06	0.18	100	0.18	100
Standard deviation	0.24	Relative SE: 0.43		Absolute SE: 0.43	
G relative 0.24					
G absolute 0.24					

[a] Curiosity (n = 55, Grand mean: 1.72, SE of the grand mean: 0.09);

[b] Decentering (n = 55, Grand mean: 1.61, SE of the grand mean: 0.11)

due to person-item and occasion-item interactions are close to zero for both TMS subscales, suggesting that the variance due to person-occasion-item interaction is mainly explained by person-occasion interaction or a state. Therefore, subsequent G-analysis is necessary to estimate the unique contribution of each variance component available in the data together with G-coefficients.

G-analysis estimates variance components and G-coefficients in both relative and absolute terms. The essential difference between them is that absolute estimates will account for all possible error variances assuming that all samples are drawn from infinite populations but relative estimates will account for finite populations in the G-study design (e.g., items). In other words, if all populations are considered as drawn from infinite populations absolute and relative variance estimates and G-coefficients will have the same values. In the current analysis (Table 3), G-coefficients are the same because error variance due to item, which is the only finite universe, is close to zero. One of possible reasons why GT has not been widely used to validate state measures is possibly because person-occasion interaction is considered as a measurement error in common G-designs with persons representing the important object of measurement. This common design was used in the current study to demonstrate its limitations and the advantages of introducing SCI to assess "stateness" of a state scale along with TCI to assess "traitness" for a measure of a trait. For instance, G-analysis (Table 3) shows error variance estimates due to different sources after accounting for the person (trait) variance. Here, error variance in both TMS subscales was mainly attributed to person-occasion interaction reflecting state changes, which is expected for a valid state measure. In the current G-analyses, person (trait) variance is assessed by G-coefficients showing values below 0.30 indicating that the TMS scores were unstable across occasions, which is consistent with expectations for a state measure. The G-analysis results mirror the traditional test-retest reliability findings and were consistent with those reported earlier for other state measures, such as a range of r values from 0.34

Table 4 Estimated person and person × occasion (P × O) interaction variance components together with G-coefficients for each individual item of the TMS subscales curiosity (*above*) and decentering (*below*)

TMS subscales and items	*P* variance	P × O variance[a]	G-coefficients[a]
Curiosity subscale			
3. Curious to learn about myself by noticing my reactions	0.06	0.51	0.11
5. Curious to see what my mind was up to from moment to moment	0.04	0.41	0.09
6. Curious about each of the thoughts and feelings I was having	0.07	0.44	0.13
10. Curious about the nature of each experience as it arose	0.09	0.43	0.17
12. Curious about my reactions to things	0.09	0.41	0.19
13. Curious to learn about myself by noticing my attention focus	0.20	0.46	0.30
Decentering subscale			
1. Experienced myself separate from thoughts and feelings	0.15	0.29	0.34
2. More concern with being open to experiences than controlling	0.12	0.52	0.19
4. Experienced my thoughts more as events than as reflection	0.00	0.49	0.00
7. Observing unpleasant thoughts and feelings without interfering	0.00	0.44	0.00
8. More invested in watching my experiences than analyzing them	0.04	0.46	0.08
9. Trying to accept each experience, pleasant or unpleasant	0.04	0.48	0.08
11. Aware of thoughts and feelings without overidentifying with them	0.03	0.35	0.07

[a] There is no difference between relative and absolute P × O variance components and G-coefficients in P × O design because there are no finite populations

to 0.46 for the State-Trait Anxiety Inventory (Ramanaiah et al. 1983; Spielberger et al. 1970, 1999). In the case of a valid trait measure, where persons (traits) explain the most variance and show stability over time, G-coefficients of 0.80 and higher would be expected (Arterberry et al. 2014).

The proposed SCI is particularly useful to assess the degree of "stateness" of a measure especially if a common G-design with persons as objects of measurement is used because person-occasion interaction (state) is treated as a measurement error in such designs. Similar to other G-estimates, the SCI was calculated based on the corrected variance components from the ANOVA (Table 2). The SCI for the curiosity subscale was 0.70 and for the decentering 0.75, which is consistent with the expectations for a valid state measure and arguably provides the first bench mark to distinguish between instruments measuring state and trait. An SCI below 0.60 would suggest that there are items in a scale, which are not sensitive to state changes (i.e., measuring a trait). In this case, modifications of an instrument should be undertaken using D-study. Similarly, the TCI can be computed to assess validity of a trait measure and modifications could be conducted if a value below 0.60 is obtained.

Besides exploration of state and trait variance components, GT analysis is also useful to identify potential sources of measurement error. In our example, the results show that error variance due to items and person-item interaction did not exceed 1%. Overall, the error variances were close to zero with the exception of interaction between person, occasion, and item in the curiosity subscale, which constitute only 8.80% with the other 92.20% explained by the state (person-occasion) component. However, both person-item and occasion-item errors were nearly zero suggesting that this error is due to state-item interaction only. If this GT method is applied to other measures, identifying sources of measurement errors can be useful especially if the values exceed 5% and hence affect the precision of a measurement. In this case, a source of measurement error (e.g., items) could be investigated in a D-study and necessary adjustments could be made to resolve the issue.

A D-study can be used to improve measurement design and to address potential issues contributing to measurement error, which is especially useful at the individual item level. Our D-study examined state and trait variance components of every individual item (Table 4) and showed that all items displayed a higher proportion of variance attributed to state compared to trait and low generalizability of scores across occasions. These findings are generally consistent with the G-study results for the complete subscales. However, items 4 and 7 in the decentering subscale showed no signs of differentiating between individual's trait levels reflected by a lack of generalizability in measuring trait. Typically, a moderate or at least a weak relationship between state and trait components is expected in a state measure (Ramanaiah et al. 1983; Spielberger 1970, 1999). Excluding those two items from the subscale was associated with a decrease in state-related variance and increase of the error variance affecting the reliability of the subscale. Therefore, items 4 and 7 were found to measure state changes only and contributed to the overall reliability of the

decentering subscale. These findings challenge the assumptions that the trait component cannot be entirely excluded in a state measure, because it is the basic predictor of a state (Hamaker et al. 2007). Assessing variance components at the individual item level could be useful because a measure may include items measuring predominantly a trait, a state, or both. In this case, state and trait items could be combined into a state and a trait subscale respectively, and neutral items excluded from the measure, which will improve accuracy in assessing state and trait.

A D-study is also useful to evaluate the appropriateness of a G-study design and the individual contribution of occasions on variability of states. For instance, removing the baseline (after holiday condition) produced a decrease of generalizability of both subscales across occasions below 0.10. This result is expected if state changes are manipulated at both occasions in the opposite direction (mindfulness exercise vs class test) and supports the appropriateness of the G-study design. Finally, attempts to optimize subscales by removing items did not yield any psychometric benefits suggesting that the TMS is an adequate measure of state mindfulness in its present form.

Limitations

The following limitations have to be acknowledged. The proposed SCI and TCI indices for validation of state and trait measures are based on the results of this study and need to be extensively tested with different instruments to establish benchmarks and cut-off points. More accurate criteria for state and trait distinctions might evolve as a result of further GT analyses of other psychometric instruments. This study was conducted with a sample of university students that has a degree of homogeneity, and the results should be replicated with larger and more diverse samples. Generalizing the results of this study (state vs trait) to the rest of the population may be limited without a truly representative sample.

In summary, the current study developed and introduced a novel and promising method to distinguish between state and trait measures using GT. The application of this method was demonstrated by generalizability analysis of the TMS—state measure of mindfulness—and provided supporting evidence for reliability and validity of the instrument. The current application of GT is recommended as an appropriate psychometric method to validate state and trait measurement tools and has the potential to open new avenues for future psychometrics work.

Acknowledgements This study is a part of doctoral work of the first author funded by the Vice-Chancellor's Scholarship of the Auckland University of Technology.

Compliance with Ethical Standards The study was conducted in compliance with the guidelines of the Auckland University of Technology Ethics Committee.

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

References

Allal, L., & Cardinet, J. (1976). *Application of generalizability theory: estimation of errors and adaptation of measurement designs*. Neuchâtel: Institut Romand de Recherche et de documentation pédagogiques.

Allen, M. J., & Yen, W. M. (1979). *Introduction on to measurement theory*. Monterey: Brooks/Cole.

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

Baer, R. (2003). Mindfulness training as a clinical intervention: a conceptual and empirical review. *Clinical Psychology: Science and Practice, 10*(2), 125–142.

Bloch, R., & Norman, G. (2012). Generalizability theory for the perplexed: a practical introduction and guide: AMEE guide no. 68. *Medical Teacher, 34*, 960–992.

Brown, K. W., & Ryan, R. M. (2003). The benefits of being present: mindfulness and its role in psychological well-being. *Journal of Personality and Social Psychology, 84*(4), 822–884.

Brennan, R. L. (1977). *Generalizability analysis: principles and procedures*. Iowa City: The American College Testing Program.

Brennan, R. L. (1992). *Elements of generalizability theory* (2nd ed.). Iowa City: ACT Publications.

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

Buss, A. H. (1989). Personality as traits. *American Psychologist, 44*, 1378–1388.

Cahn, B. R., & Polich, J. (2006). Meditation states and traits : EEG, ERP, and neuroimaging studies. *Psychological Bulletin, 132*, 180–211.

Chaplin, W. F., John, O. P., & Goldberg, L. R. (1988). Conceptions of states and traits: dimensional attributes with ideals as prototypes. *Journal of Personality and Social Psychology, 54*(4), 541–557.

Chiesa, A., & Serretti, A. (2010). A systematic review of neurobiological and clinical features of mindfulness meditations. *Psychological Medicine, 40*, 1239–1252.

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

Davis, K. M., Lau, M. A., & Cairns, D. R. (2009). Development and preliminary validation of a trait version of the Toronto Mindfulness Scale. *Journal of Cognitive Psychotherapy, 23*(3), 185–197.

Epstein, S. (1984). Trait theory as personality theory: can a part be as great as the whole? *Psychological Inquiry, 5*, 120–122.

Fisher, R. A. (1925). *Intraclass correlation and the analysis of variance. In: statistical methods for research workers*. New Delhi: Cosmo Publications for Genesis Pub.

Gardinet, J., Johnson, S., & Pini, G. (2009). *Applying generalizability theory using EduG*. New York: Routledge.

Geiser, C., Litson, K., Bishop, J., Keller, B. T., Burns, G. L., & Servera, M. (2015). *Analyzing person, situation and person - situation interaction effects: Latent State-Trait Models for the Combination of Random and Fixed Situations*.

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

Hofmann, S. G., Sawyer, A. T., Witt, A., & Oh, D. (2010). The effect of mindfulness-based therapy on anxiety and depression: a meta-analytic review. *Journal of Consulting and Clinical Psychology, 78*(2), 169–183.

Ivanovski, B., & Malhi, G. S. (2007). The psychological and neurophysiological concomitants of mindfulness forms of meditation. *Acta Neuropsychiatrica, 19*, 76–91.

Kenny, D. A., & Zautra, A. (Eds.). (2001). *Trait–state models for longitudinal data*. Washington DC: American Psychological Association.

Lau, M. A., Bishop, S. R., Segal, Z. V., Buis, T., Anderson, N. D., Carlson, L., et al. (2006). The Toronto mindfulness scale: development and validation. *Journal of Clinical Psychology, 62*(12), 1445–1467.

Park, T., Reilly-Spong, M., & Gross, C. R. (2013). Mindfulness: a systematic review of instruments to measure an emergent patient-reported outcome (PRO). *Quality of Life Research, 22*, 2639–2659.

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

Rosenzweig, S., Greeson, J. M., Reibel, D. K., Green, J. S., Jasser, S. A., & Beasley, D. (2010). Mindfulness-based stress reduction for chronic pain conditions: variation in treatment outcomes and role of home meditation practice. *Journal of Psychosomatic Research, 68*(1), 29–36.

Segal, Z. V., Williams, J. M. G., & Teasdale, J. D. (2013). *Mindfulness-based cognitive therapy for depression* (2nd ed.). New York: Guilford Press.

Shavelson, R. G., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist, 44*, 599–612.

Shoukri, M. M., Asyali, M. H., & Donner, A. (2004). Sample size requirements for the design of reliability study: review and new results. *Statistical Methods in Medical Research, 13*, 251–271.

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

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

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

Williams, M., & Penman, D. (2011). *Mindfulness: an eight-week plan to find peace in a frantic world*. New York: Rodale Inc..