International Psychogeriatrics (2022), 34:12, 1023–1033 © International Psychogeriatric Association 2021
doi:10.1017/S1041610221000363

Applying generalizability theory to examine assessments of subjective cognitive complaints: whose reports should we rely on – participant versus informant?

Q. C. Truong,[1] C. Choo,[1] K. Numbers,[2] A. G. Merkin,[3,4] H. Brodaty,[2] N. A. Kochan,[2] P. S. Sachdev,[2] V. L. Feigin,[3] and O. N. Medvedev[1]

[1]*School of Psychology, University of Waikato, Hamilton, New Zealand*
[2]*Centre for Healthy Brain Ageing, School of Psychiatry, Faculty of Medicine, University of New South Wales, Sydney, Australia*
[3]*Auckland University of Technology, Auckland, New Zealand*
[4]*Centre for Precise Psychiatry and Neurosciences, Germany/Switzerland*

ABSTRACT

Objectives: This study aimed to apply the generalizability theory (G-theory) to investigate dynamic and enduring patterns of subjective cognitive complaints (SCC), and reliability of two widely used SCC assessment tools.

Design: G-theory was applied to assessment scales using longitudinal measurement design with five assessments spanning 10 years of follow-up.

Setting: Community-dwelling older adults aged 70–90 years and their informants, living in Sydney, Australia, participated in the longitudinal Sydney Memory and Ageing Study.

Participants: The sample included 232 participants aged 70 years and older, and 232 associated informants. Participants were predominantly White Europeans (97.8%). The sample of informants included 76 males (32.8%), 153 females (65.9%), and their age ranged from 27 to 86 years, with a mean age of 61.3 years (SD = 14.38).

Measurements: The Memory Complaint Questionnaire (MAC-Q) and the Informant Questionnaire on Cognitive Decline in the Elderly (IQCODE).

Results: The IQCODE demonstrated strong reliability in measuring enduring patterns of SCC with G = 0.86. Marginally acceptable reliability of the 6-item MAC-Q (G = 0.77–0.80) was optimized by removing one item resulting in G = 0.80–0.81. Most items of both assessments were measuring enduring SCC with exception of one dynamic MAC-Q item. The IQCODE significantly predicted global cognition scores and risk of dementia incident across all occasions, while MAC-Q scores were only significant predictors on some occasions.

Conclusions: While both informants' (IQCODE) and self-reported (MAC-Q) SCC scores were generalizable across sample population and occasions, self-reported (MAC-Q) scores may be less accurate in predicting cognitive ability and diagnosis of each individual.

Key words: aging, cognitive assessment, dementia, mental capacity, generalizability theory, longitudinal design

Introduction

Older adults often report subjective cognitive complaints (SCC), which relate to an individual's self-experience of cognitive deterioration (Hildreth and Church, 2015). Currently, SCC contribute to the criteria for a diagnosis of mild cognitive impairment (MCI) (Petersen, 2016; Winblad *et al.*, 2004) and may be considered as the earliest detectable stage of preclinical dementia (Jonker *et al.*, 2000; Mitchell *et al.*, 2014). SCC can be self-reported or reported by informants (e.g. family member or friend) with the advantage of capturing daily cognitive and memory changes that standardized neuropsychological tests may not detect (Brodaty *et al.*, 2002; Jorm *et al.*, 1991; Numbers *et al.*, 2020).

Despite the potential benefits of SCC assessments, it remains questionable as to whether self-reported

Correspondence should be addressed to: Oleg N. Medvedev, School of Psychology, University of Waikato, Private Bag 3105, Hamilton 3240, New Zealand. Phone: + 64 7 837 9212. Email: oleg.medvedev@waikato.ac.nz Received 17 Dec 2020; revision requested 12 Feb 2021; revised version received 23 Feb 2021; accepted 27 Feb 2021. First published online 13 April 2021.

SCC reliably predict objective cognitive performance and/or dementia incident. A relationship between self-reported SCC and cognitive impairment ranges from negligible (e.g. Burmester *et al.*, 2016) to none (e.g. Lenehan *et al.*, 2012; Reid and MacLullich, 2006). One explanation for such inconsistency may be the influence of mood and certain personality traits on complaining behaviors (Ponds and Jolles, 1996). It is well established that subjective impressions of decline are exacerbated by depression and anxiety as well as personality traits such as neuroticism and conscientiousness (Reid and MacLullich, 2006). Therefore, subjective reports of cognitive ability provided by close informants may present more reliable approximation of objective cognitive performance (Slavin *et al.*, 2015) and future cognitive decline (Caselli *et al.*, 2014). Furthermore, in the clinical setting, informant-reported SCC are often increasingly relied upon as individuals progress through preclinical stages of dementia and begin losing insight into their cognitive changes over the debilitating course of dementia (American Psychiatric Association, 1994). However, no empirical examination was conducted to date using an appropriate methodology to investigate whose reports (i.e. participants' or informants') are more reliable, and at what stage researchers and clinicians should rely on which reports. As dementia is typically marked by insidious onset and gradual progression, a longitudinal design will be ideal in tracking any cognitive change over time (American Psychiatric Association, 1994).

Moreover, it is important to differentiate reliably between dynamic and enduring SCC patterns over longer time. A reliable trait measure would reflect enduring changes over time (e.g. alterations in long-term subjective cognition) and remain unaffected by individual's transient changes (e.g. mood or current stress level). Conversely, a state measure would be sensitive to dynamic changes, which may confound assessment of long-term subjective cognition. While the widely used SCC measures such as the Memory Complaint Questionnaire (MAC-Q) (Crook *et al.*, 1992) and the Informant Questionnaire on Cognitive Decline in the Elderly (IQCODE) (Jorm, 1994) have good internal consistency, it does not support their ability to distinguish between enduring and dynamic patterns of SCC. Internal consistency coefficients (e.g. Cronbach's alpha) are not appropriate to estimate temporal reliability of scales because they only estimate consistency or inter-correlations between individual items at one time point. Moreover, test–retest reliability coefficients often used to distinguish between dynamic and enduring patterns have limited accuracy because these are merely correlations between total scale scores at two different times (e.g. Time 1 and Time 2). For example, if a person improves on one symptom but gets worse on another, the total score remains the same without reflecting clinically important changes. Therefore, these coefficients do not account for variability of individual items over time and other sources of measurement error such as the effects of item, occasion, person, and their interactions (Bloch and Norman, 2012; Medvedev *et al.*, 2020). For instance, a response to an item may depend on assessment occasion rather than changes in individual's performance. A comprehensive estimation of reliability is therefore required and generalizability theory (G-theory) was advocated as the most appropriate method to investigate dynamic and enduring patterns in a measure and examine reliability and generalizability of assessment scores (Medvedev *et al.*, 2017; Truong *et al.*, 2020).

G-theory is a successor of classical test theory (CTT) and is particularly well suited to examine the overall reliability of psychometric instruments (Brennan, 2010; Shavelson and Webb, 1991). While CTT postulates that any measurement consists of true variance and error variance presented as a single factor, G-theory utilizes ANOVA to estimate all possible sources of error variance that may affect the main outcome variable as well as the accuracy of the measurement itself (Allen and Yen, 2001; Cronbach *et al.*, 1963). Furthermore, CTT evaluates the reliability of a measure at only one aspect (e.g. internal consistency) at a time or examines the distinction between dynamic and enduring patterns of a measure using test–retest reliability coefficients. G-theory extends CTT and simultaneously examines all potential sources of error variance that may influence reliability such as person, scale items, occasion, and all their interactions (Medvedev *et al.*, 2017; Shavelson *et al.*, 1989). Many studies have demonstrated applicability of G-theory as the most appropriate method for estimating the overall reliability and generalizability of assessment scores and distinction between dynamic and enduring patterns in a measure (Arterberry *et al.*, 2014; Medvedev *et al.*, 2017; Paterson *et al.*, 2018; Truong *et al.*, 2020). Therefore, applying G-theory can be useful to examine and improve the precision of a psychometric instrument as well as to differentiate between enduring and dynamic patterns reflected by such measure.

The aim of the current study was to apply G-theory to examine reliability and distinguish between dynamic and enduring patterns in the self-report MAC-Q and informants IQCODE SCC assessment tools. A longitudinal design was utilized with participants assessed at five occasions, separated by 2–4 years intervals. Application of G-theory involved two parts: a generalizability study (G-study) and a decision study (D-study). The

https://doi.org/10.1017/S1041610221000363 Published online by Cambridge University Press

purpose of the G-study was to examine the overall generalizability of the MAC-Q and IQCODE and evaluate sources of error variance in each measure. D-study aimed to subsequently evaluate psychometric properties of individual items of these two scales and to manipulate measurement design to optimize the reliability of measurement (Cardinet *et al.*, 2011; Shavelson *et al.*, 1989). We also aimed to evaluate the utility of these scales for predicting incident dementia and global cognition scores.

Method

Participants

Community-dwelling older adults aged 70–90 years, living in the Eastern Suburbs of Sydney, Australia, were selected via the electoral roll and invited to participate in the Sydney Memory and Ageing Study (MAS) (Sachdev *et al.*, 2010). Of 8,914 individuals invited to participate, 1,037 participants were included in the baseline sample (occasion 1). Inclusion criteria were the ability to speak and write English sufficiently well to complete a psychometric assessment and self-report questionnaires. Exclusion criteria were any major psychiatric diagnoses, acute psychotic symptoms, current diagnosis of multiple sclerosis, motor neuron disease, developmental disability, progressive malignancy, and/or dementia. All participants provided written consent to participate in this study, which was approved by the University of New South Wales Human Ethics Review Committee (HC 05037, 09382, 14327). More detailed methods of recruitment and baseline demographics have been previously described by Sachdev and colleagues (Sachdev *et al.*, 2010). Of the 1037 participants included in the present study, 1,009 (97.3%) had an informant. Informants were selected by nominations of participants. Informants answered questions relating to the participant's memory, thinking, and daily functioning. Qualified informants were those who had at least 1 hour of contact with the participant per week; on average, they had 8.3 hours of weekly contact. All participants and informants provided written consent to participate in this study, which was approved by the University of New South Wales Human Ethics Review Committee (HC 05037, 09382, 14327).

Figure 1 presents consort diagram including the number of MAC-Q and IQCODE reports, and the number of participants diagnosed with dementia along with computed their global cognition scores, for each occasion (wave). Of the 1,009 participants with informants, 232 (23%) had reports of MAC-Q and IQCODE at all five occasions and were included in the G-analyses. We excluded participants' with informants (77%) whose MAC-Q and IQCODE were incomplete at one or more occasions. The MAC-Q or IQCODE data were missing at some waves because either the participant or informant was not contactable or was not able to do the assessment at that wave. In some instances, participants were too ill or advanced in dementia to answer questions in later waves (informant only), and in others, participants simply did not have an informant who was willing to complete an interview or questionnaire on their behalf. The ethnicity of the extracted sample was predominantly White Europeans (97.8%); the remaining sample was 0.4% other and 1.7 % unrevealed. Informants from the extracted sample were 76 males (32.8%), 153 females (65.9%), and their age ranged from 27 to 86 years, with a mean age of 61.3 years (SD = 14.38). Missing responses per item of either the MAC-Q or the IQCODE in the extracted sample comprised less than 0.05% which were negligible and thus substituted by mean imputation at each respective wave (Huisman, 2000). This sample size of 232 participants exceeded the required sample size of 84 participants for repeated measures ANOVA over five occasions needed to accomplish the power (1-β) of 0.95 to detect effect size of 0.15 under p value of 0.05.

Measures

The MAC-Q (Crook *et al.*, 1992) is a well-validated unidimensional 6-item questionnaire. Internal consistency Cronbach's alpha of the MAC-Q was reported in the range from 0.57 to 0.88 with most studies indicating acceptable values and confirming unidimensionality of the scale (Buckley *et al.*, 2013; Crook *et al.*, 1992; Reid *et al.*, 2012). The MAC-Q asked participants to rate themselves compared to how they previously performed on several everyday memory tasks (e.g. difficulty remembering names; see Supplemental Materials for a full list of items). At occasion 1, participants received the conventional MAC-Q wording "How would you rate yourself compared to 5 years ago," but for each subsequent occasion, the wording was changed to "How would you rate yourself compared to 2 years ago" to capture the intervening time between assessments. Participants rated themselves for each item on a scale of 1 to 5; total score range from 5 to 30, with higher scores indicating greater subjective memory loss.

The IQCODE (Jorm, 1994) consists of 16 items that asks informants to report on their perceived changes of the participant's cognition and functioning. Each item is scored on a 5-point Likert scale with options ranging from 1 = "much improved" to 5 = "much worse." The IQCODE is completed by informants who are well known to the individual (Harrison *et al.*, 2016) and has been shown to reliably predict incident dementia (Numbers *et al.*, 2020). The original IQCODE consists of 26

https://doi.org/10.1017/S1041610221000363 Published online by Cambridge University Press

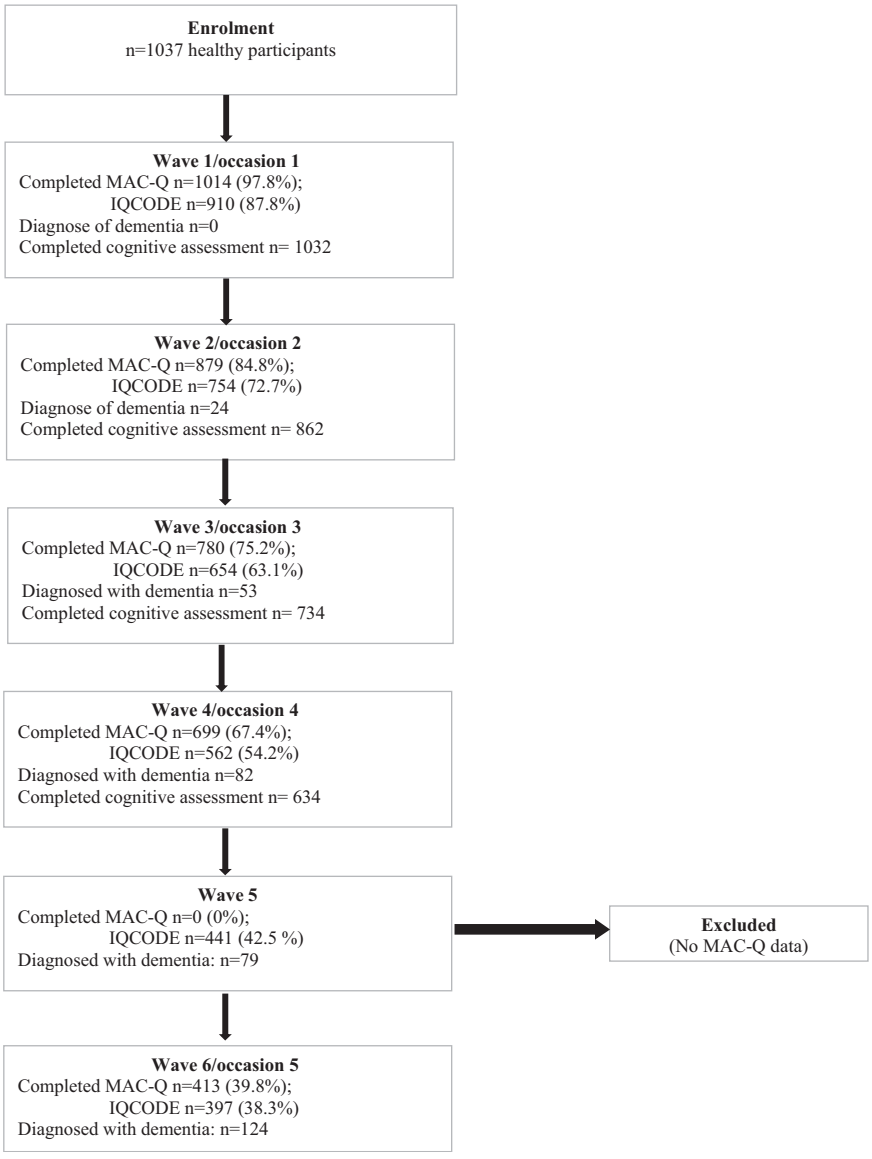

Figure 1. CONSORT diagram for participants who completed cognitive assessments inclusive of dementia-diagnosed cases at each wave/occasion.

questions/items. An abbreviated version of the IQCODE consisting of 16 items has been found to perform as reliably as the original version (Jorm, 1994, 2004), with a number of studies confirming high internal consistency (Cronbach's alpha = 0.93 to 0.97) (Harrison *et al.*, 2015; Phung *et al.*, 2015; Tang *et al.*, 2004) and a superior ability to predict incident dementia (Park, 2017; Perroco *et al.*, 2008).

Dementia diagnosis

Clinical diagnoses were performed for all occasions (10-year follow-up). At occasion 1, and at each 2-year follow-up, individuals were brought to a consensus review meeting where at least three clinicians from a panel of neuropsychiatrists, psychogeriatricians, and neuropsychologists discussed all available clinical, neuropsychological, laboratory, and imaging data to reach a consensus diagnosis. A diagnosis of dementia was based on the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV) (American Psychology Association, 1994), that is, the development of one or more cognitive deficit(s) that represent a decline from a previous level of performance and were sufficiently severe as to cause impairment in functioning (the Bayer Activities of Daily Living Scale [B-ADL] score $\geq$ 3.0). Individuals who did not receive a dementia diagnosis were classified as "not dementia" at each occasion, and no dementia cases were present at occasion 1 as this was an exclusionary criterion (Sachdev *et al.*, 2010).

https://doi.org/10.1017/S1041610221000363 Published online by Cambridge University Press

Objective cognitive performance

Comprehensive cognitive data were available for occasion 1 to occasion 4 (6-year follow-up), only. Cognitive performance over these first four occasions was assessed using a comprehensive neuropsychological test battery which comprised 10 tests that measured the domains of attention/processing speed, language, executive function, visuospatial ability, and memory (see Supplementary Table S1). Domain and global cognition composites were computed as standardized z-scores as follows. First, raw test scores were converted to z-scores using the baseline means and standard deviations (SDs) of a reference group which comprised 732 MAS participants classified as cognitively healthy at occasion 1 (native English speakers with a Mini-Mental State Examination score of 24 or above, no evidence of dementia or current depression, no history of delusions or hallucinations, and no major neurological disease, significant head injuries, progressive malignancies or central nervous system (CNS) medications). Of the 732 participants (ages ranged from 70.29 to 90.80 years with M = 78.57, SD = 4.72), 219 (29.9%) completed tertiary qualification, 128 (17.5%) completed high school and/or diploma, 350 (47.8%) were not completed high school/diploma, 23 (3.2%) were not completed tertiary qualification, and 12 (1.6%) completed primary school. Secondly, composite domain scores were formed by averaging the z-scores of the component tests (as defined above), apart from the visuospatial domain which was represented by a single test. Each domain composite was standardized by transforming, so that the mean and SD of the baseline cognitively healthy group were 0 and 1, respectively. Finally, global cognition scores for each occasion were calculated by averaging the domain z-scores and again transforming these scores so that the means and SDs for the baseline reference group were 0 and 1, respectively.

Data analyses

EduG 6.1-e software (Swiss Society for Research in Education Working Group, 2006) was used to conduct generalizability analyses by following the guidelines described in Truong *et al.* (2020). Both G-study and D-study used two-facet design (person by item by occasion) which item (I) and occasion (O) were two facets of interest (instrumentation facets) and person (P) was the object of measurement (differentiation facet), expressed as P x I x O (Cardinet *et al.*, 2011; Truong *et al.*, 2020; Vispoel *et al.*, 2018). The facet of I was fixed because the same items of assessments were used across all participants and all occasions, whereas the P and O facets were infinite. Besides that, the facet P was not a source of error and in a study employing G-theory method, all error variances are counted as 100% after controlling for person variance (P), which reflects true differences between persons (Cardinet *et al.*, 2011). G-theory estimates for the design of person by item by occasion expressed as P x I x O were calculated using formulae included in Supplementary Table S2 (Shavelson *et al.*, 1989).

There are two reliability coefficients, relative G-coefficient (G_r) and absolute G-coefficient (G_a), for the object of measurement (person) in a generalizability study. The relative model of measurement is based on a norm-referenced manner in which a person's assessment score is compared against the scores of others (Vispoel *et al.*, 2018). G_r accounts for a relative error variance (σ_δ^2) which is related to the I facet (object of measurement) that may affect a relative measurement (e.g. interaction between person and occasion – PxQ, and interaction between person and item – PxI) and includes divisions by desired sample sizes (Shavelson and Webb, 1991). Both G coefficients are estimating reliability of an enduring pattern of a measurement if the person (P) is differentiation facet. Specifically, G_r of 0.80 or higher is determined as good reliability of assessment score (Cardinet *et al.*, 2011), while Ga above 0.70 is considered as acceptable reliability (Arterberry *et al.*, 2014; Truong *et al.*, 2020). Both a state component index (SCI) and trait component index (TCI) were obtained, which represent the variance proportion attributed to a dynamic (state) and an enduring (trait) pattern in a measure (Medvedev *et al.*, 2017). SCI of 0.60 or higher (TCI < 0.40) would indicate that variance is reflecting a dynamic pattern. On the contrary, TCI above 0.60 (SCI < 0.40) would signify that variance is reflecting an enduring pattern. In the D-study, variance components were computed for each individual item, and effects of removing facets levels were examined to optimize the reliability of the MAC-Q and the IQCODE.

IBM SPSS Statistics 25 software was used to compute estimates that related to CTT approaches and descriptive statistics including means, SD, Cronbach's alpha, and intraclass correlation coefficient (ICC) for the IQCODE and the MAC-Q. Logistic regression analyses were conducted to examine how the IQCODE and the MAC-Q were able to predict the incidence of dementia across occasions 2–5. Three logistic regression models were carried out for each occasion. Each logistic regression model involved the outcome variable of dementia diagnosed with one predictor of either MAC-Q scores or IQCODE scores at the same occasion. Moreover, linear regression analyses were also used to estimate ability of these measures to predict global cognition scores across the first four

https://doi.org/10.1017/S1041610221000363 Published online by Cambridge University Press

occasions. Three independent linear regression models were conducted at each occasion with the outcome variable of global cognition scores and either MAC-Q scores or IQCODE scores as a predictor. Prior to all regression analyses, assumption tests were conducted to screen for potential violations.

Results

G-study

The variance components attributed to person (P), item (I), and occasion (O), and their interactions (PxI, PxO, IxO, and PxIxO) together with generalizability coefficients and state and trait component indices are presented in Table 1 for the MAC-Q and the IQCODE. The IQCODE showed better reliability and generalizability of scores across persons and occasions, with both relative and absolute G coefficients of 0.86, 95% CI [0.84; 0.88]. Measurement error was predominantly explained by PxI and PxO interactions for the IQCODE, which together explained 77.1% of the total error variance after accounting for the true person variance. Slightly lower, but still acceptable values (Gr = 0.80, 95% CI [0.77; 0.83]; Ga = 0.77, 95% CI [0.73; 0.81]) were observed for the 6-item MAC-Q, with the main source of error variance due to the PxI interaction explaining 35.7% of the total error variance. Consistent with reliability estimates, TCI values were 0.96 for the MAC-Q and 0.95 for the IQCODE indicating that both instruments reliably assess enduring patterns of SCC.

Table 1. G-study estimates for the MAC-Q and IQCODE including standard errors of the grand mean (SE), coefficient G relative (Gr), coefficient G absolute (Ga), trait component index (TCI), state component index (SCI), grand mean (GM), variance components (%), and for the Person (P) × Occasion (O) × Item (I) design including interactions (n = 232)

	MAC-Q		IQCODE	
FACETS	σ^2	%	σ^2	%
P	0.044		0.019	
I	0.000	2.9	0.000	2.2
O	0.001	5.5	0.000	2.1
PxI	0.005	35.7	0.001	39.6
PxO	0.002	12.6	0.001	37.5
IxO	0.000	3.5	0.000	1.5
PxIxO	0.005	39.7	0.001	17.1
Grand mean	3.224		3.155	
SE	0.042		0.137	
Gr	**0.80**		**0.86**	
Ga	**0.77**		**0.86**	
TCI	0.96		0.95	
SCI	0.04		0.05	

Note: Numbers in bold signify acceptable reliability/generalizability coefficients.

D-study

A series of generalizability analysis were conducted to obtain variance components for each individual items of the MAC-Q and IQCODE. The item-level estimates for variance of person, occasion, and person–occasion interaction, together with computed SCI, are included in Supplementary Table S3. There was only one MAC-Q item reflecting high sensitivity for transitory changes in SCC patterns over time; "item-e: Remembering the item[s] you intended to buy when you arrive at the supermarket store or pharmacy?", which had the highest SCI of 0.66. The other five items of the MAC-Q revealed an SCI range from 0.15 to 0.45 indicating a lower proportion of variance associated with dynamic changes in SCC over time. However, all IQCODE items reflected predominantly enduring patterns of SCC.

Six additional generalizability analyses were conducted by excluding one item at a time for the MAC-Q, as we expected that this may result in improving the reliability of the scale in measuring enduring patterns of SCC (Supplementary Table S4a). The first analysis involved removing the first item (item-a), with subsequent analyses removing one item at a time and examining reliability. Removing the final item (f) of the MAC-Q "In general, how would you describe your memory as compared to 10 years ago?" was the only analysis that resulted in improvement of both relative and absolute G coefficients above 0.80 benchmark, suggesting that the 5-item MAC-Q (i.e. MAC-Q without item f) has better reliability compared to the 6-item MAC-Q. Next analyses involved removing one occasion at a time for the MAC-Q (Supplementary Table S4b) to examine how this affects the reliability of the scale. Removing any occasion only slightly decreased both G coefficients, which remained in the acceptable range.

Additional G-analyses were conducted on the IQCODE (Supplementary Table S5), which involved removing items more sensitive to dynamic SCC with SCI ≥ 0.40 in attempt to optimize reliability. This resulted in lower G-coefficients compared to the original IQCODE. Additionally, removing one occasion at a time for the IQCODE only slightly decreased G-coefficients. These findings together support reliability of the IQCODE with the current measurement design.

CTT analyses

Descriptive statistics for the 5-item MAC-Q, the 6-item MAC-Q, and the IQCODE at five occasions are presented in Supplementary Table S6. The

https://doi.org/10.1017/S1041610221000363 Published online by Cambridge University Press

Table 2. Logistic regression model coefficients for the MAC-Q and IQCODE variables across occasions 2–5 predicting the incidence of dementia

Predicting diagnosis	MPC	β	SE (β)	P	Exp(β) [95% CI]
Occasion 2:					
MAC-Q (5-item)	97.6	1.21	0.13	0.35	1.13 [0.88, 1.45]
MAC-Q (6-item)	96.9	−0.10	0.11	0.37	0.91 [0.73, 1.13]
IQCODE	98.2	5.73	0.90	<0.001	308.80 [53.24, 1791.26]
Occasion 3					
MAC-Q (5-item)	93.9	0.28	0.07	<0.001	1.32 [1.14, 1.53]
MAC-Q (6-item)	94.0	0.30	0.06	<0.001	1.35 [1.19, 1.53]
IQCODE	94.8	4.05	0.53	<0.001	57.13 [20.07,162.66]
Occasion 4					
MAC-Q (5-item)	90.8	0.29	0.06	<0.001	1.34 [1.20, 1.50]
MAC-Q (6-item)	90.9	0.25	0.05	<0.001	1.29 [1.17, 1.42]
IQCODE	92.3	3.58	0.39	<0.001	36.01 [16.74, 77.48]
Occasion 5					
MAC-Q (5-item)	80.8	0.16	0.07	0.02	1.17 [1.03, 1.33]
MAC-Q (6-item)	78.2	0.05	0.30	0.30	1.05 [0.96, 1.14]
IQCODE	85.3	3.65	0.41	<0.001	38.60 [17.31, 86.108]

Note: MPC: model percentage correct; Exp(β): the exponentiation of the β-coefficient.

internal consistency Cronbach's alpha of the 6-item MAC-Q was fair to good over five occasions and ranged between 0.57 and 0.76, which was consistent with values reported by other studies (Buckley *et al.*, 2013, Crook *et al.*, 1992; Reid *et al.*, 2012). Given lower Cronbach alpha, we explored a possibility of multidimensionality that could impact on internal consistency of the MAC-Q with the current dataset using full sample at occasion 1 (n = 1011) by applying exploratory factor analysis. The results showed only one factor with eigenvalue >1 using Kaiser criterion, and the scree plot clearly indicated elbow after one factor with loading on the first principle component ranging from 0.53 to 0.75 supporting unidimensionality of the MAC-Q. Temporal stability was supported by ICC of 0.84 across all occasions. The mean scores of the 5-item MAC-Q were only significantly different between occasion 1 and 2, while that of the 6-item MAC-Q were significantly different between occasion 1 and occasions 2 and 3. The IQCODE demonstrated higher internal consistency with Cronbach's alphas ranging from 0.84 to 0.95, though the ICC of 0.70 was lower than both MAC-Q scales. Overall, the MAC-Q and IQCODE scales showed acceptable to high internal reliability and acceptable temporal reliability for a measure of enduring patterns over time, which is consistent with G-study results.

To evaluate predictive validity of the original MAC-Q (6-item), the shortened MAC-Q (5-item), and the IQCODE, a series of binary logistic regression analyses were conducted to predict risk of incident dementia. Table 2 presents coefficients for the models' predictors as well as model percentage correct, for the three SCC measures across the four follow-up occasions (occasion 1 was excluded because all participants were initially healthy). Prior to the analyses, assumption testing was conducted for all the models and did not indicate any violations. The Hosmoer and Lemeshow tests indicated good fit for these logistic regression models (all p's > 0.05). Accuracy of all models across occasions were ranging from 78.2% to 98.2% in their predictions of incidence of dementia. The IQCODE significantly predicted incident dementia at all examined occasions, with all p's < 0.001. Whereas, the two versions of the MAC-Q (i.e. 5-item vs. 6-item) were only significant predictors of dementia incident on occasions 3 and 4 with p's $\leq$ 0.02 but not at occasions 2 and 5 (p's $\geq$ 0.30).

Table 3 presents a series of linear regression analyses conducted to determine the relationship between predictors such as the two MAC-Q scales and the IQCODE, and the outcome measured as participants' global cognition scores at the first four occasions. Several assumptions were evaluated for these linear regression models prior to the interpretation of the results. The data were distributed close to normal with skewness values for all variables in the models ranging from − 1.03 (dementia diagnose at occasion 4) to 2.12 (5-item MAC-Q scores at occasion 5). Inspection of normal probability plots of regression standardized residuals also indicated that these variables were normally distributed. The scatterplots of standardized residuals were compared against standardized predicted values and also revealed that these variables met the assumptions of linearity and homoscedasticity of residuals and were free from univariate outliers. Finally, there were no multicollinearity issue because only one predictor was independently tested for each model.

https://doi.org/10.1017/S1041610221000363 Published online by Cambridge University Press

Table 3. Linear regression model coefficients for the MAC-Q and IQCODE variables across occasions 1–4 predicting the global cognition scores

PREDICTING COGNITION	R^2	β	SE (β)	P	STANDARDIZED β [95% CI]
Occasion 1					
MAC-Q (5-item)	0.002	− 0.04	0.03	0.13	− 0.05 [− 0.11, 0.01]
MAC-Q (6-item)	0.004	− 0.04	0.02	0.06	− 0.06 [0.10, 1.16]
IQCODE	0.030	− 1.02	0.19	<0.001	− 0.17 [− 0.24, − 0.11]
Occasion 2					
MAC-Q (5-item)	0.022	− 0.14	0.03	<0.001	− 0.15 [− 0.22, − 0.08]
MAC-Q (6-item)	0.012	− 0.08	0.03	<0.01	− 0.11 [− 0.18, − 0.04]
IQCODE	0.055	− 1.31	0.20	<0.001	− 0.24 [− 0.31, − 0.17]
Occasion 3					
MAC-Q (5-item)	0.011	− 0.10	0.04	<0.01	− 0.11 [− 0.18, − 0.03]
MAC-Q (6-item)	0.014	− 0.09	0.03	0.00	− 0.12 [− 0.18, − 0.05]
IQCODE	0.040	− 0.94	0.19	<0.001	− 0.20 [− 0.28, − 0.12]
Occasion 4					
MAC-Q (5-item)	0.050	− 0.21	0.04	<0.001	− 0.23 [− 0.31, − 0.15]
MAC-Q (6-item)	0.047	− 0.16	0.03	0.00	− 0.22 [− 0.29, − 0.14]
IQCODE	0.085	− 1.12	0.16	<0.001	− 0.29 [− 0.37, − 0.21]

Note: SE (β): standard error of the β-coefficient.

The IQCODE significantly predicted global cognition scores for all four occasions, with all p's < 0.01, while both MAC-Q versions were not significantly associated with global cognition at the first occasion (p's > 0.05)."

Discussion

The aim of the current study was to evaluate and optimize reliability and distinguish between dynamic and enduring patterns in the MAC-Q (self-report) and IQCODE (informants) SCC measures using G-theory. The results showed that the optimized 5-item MAC-Q and the IQCODE were reliable in measuring enduring pattern of SCC with G-coefficients of 0.80 and higher, and index of trait (TCI) above 0.94, suggesting that their scores are generalizable across sample population and occasions. In line with previous research (Slavin *et al.*, 2015), we found that the IQCODE SCC scores significantly predicted risk of dementia incident and global cognition across all occasions, while the MAC-Q scores were only significant predictors on some occasions. However, these results should be interpreted with caution due to differences in length and format between the IQCODE and the MAC-Q. Together our findings suggest that the MAC-Q reliably measures individual levels of SCC, but these self-reported SCC may be less accurate in reflecting cognitive abilities and diagnosis of each individual. It is possible that the MAC-Q tends to reflect individual tendencies to report complaints (e.g. about their self-perceived memory errors) rather than their actual cognitive capacities. In other words, some people may have stronger tendency to ruminate on everyday memory errors or lean toward complaining behavior, which may not necessary reflect their actual cognitive abilities. Consistent with the recent clinical literature, SCC may be related to anxiety and stress in individuals with normal cognition (Chin *et al.*, 2019). The outcome of our study has clinical implications, which underscore the importance for clinicians to seek corroboratory evidence from knowledgeable informants in their follow-up of aging patients. This, in turn, could help in detection of MCI, which is the preclinical stage of the trajectory of cognitive decline, and would assist in ongoing clinical management and planning, as once dementia is diagnosed, it runs a debilitating course (Langa and Levine, 2014).

A D-study was conducted to examine psychometric properties of individual items of the MAC-Q and IQCODE in an effort to optimize reliability of the measurement. Results showed that most individual items of the IQCODE and MAC-Q measured enduring patterns of SCC, except item-e of the MAC-Q (item-e: "Remembering the item[s] you intended to buy when you arrive at the supermarket store or pharmacy"). However, removing this item did not improve the reliability of the MAC-Q in measuring enduring pattern of SCC. Similar results were found when removing each item of the MAC-Q one at a time, with the exception of the last item (item-f: "In general, how would you describe your memory as compared to 10 years ago?"). Removing this last item boosted the marginally acceptable reliability of the 6-item MAC-Q (G = 0.77–0.80) up to G = 0.80–0.81 in the optimized 5-item MAC-Q. No reliability

https://doi.org/10.1017/S1041610221000363 Published online by Cambridge University Press

improvements were achieved by manipulating measurement design of the IQCODE, suggesting optimal reliability of the scale in the current measurement design.

Strengths and limitations

The main strength of the study was to apply the comprehensive methodology of G-theory to a relatively large sample in a longitudinal study spanning over 10 years. However, limitations need to be acknowledged. The study was conducted with participants recruited from a relatively small catchment area in Sydney, Australia. Moreover, the participants belonged to a predominantly White (European) ethnic group, and the generalizability to other ethnicities is questionable. Recent research suggests that cultural variations contribute to vulnerabilities and resilience across a range of health issues (Choo *et al.*, 2017). As such, it would be beneficial to replicate these analyses on samples comprising other ethnicities, including culturally and linguistically diverse groups. This study aimed to analyze data from five occasions with equal intervals of 2 years; however, the interval between occasion 4 and occasion 5 was 4 years, as MAC-Q data were missing for wave 5 (8-year follow-up) assessments. Future studies should endeavor to replicate these analyses using equal intervals between occasions.

The findings of this study added evidence supporting the benefits of using the informant SCC report. However, due to differences between the IQCODE and the MAC-Q format, more accurate comparison between informants' reports measured by the IQCODE and self-reports could be achieved using the self-report version of the IQCODE – the Informant Questionnaire on Cognitive Decline-Self-report (IQCODE-SR) (Jansen *et al.*, 2008). Therefore, further studies are warranted to compare the IQCODE with the IQCODE-SR.

In addition, this study did not control for demographic variables (e.g. age, sex, socioeconomic, and education status), mood, or personality in both sets of regression analyses. Nevertheless, the results of G-analyses indicated that the IQCODE and both versions of MAC-Q were measuring enduring pattern of SCC and were less affected by dynamic and transient conditions such as mood. Notably, less than 20% of variance was explained by error due to temporal factor and interactions ($G = 0.80–0.86$).

Conclusion

The findings of this study indicated that the IQCODE and MAC-Q assessment scores were generalizable across sample population and occasions and captured enduring patterns of SCC over 10 years. The optimized 5-item MAC-Q was superior to the original 6-item scale when assessing SCC over time. While clinicians and researchers could rely on both participants and informants' SCC reports of the IQCODE and the MAC-Q, self-reported (MAC-Q) scores may be less accurate in predicting cognitive ability and diagnosis of each individual.

Compliance with ethical standards

The study complied with the guidelines of the University Ethics Committee, which are consistent with internationally accepted ethical standards.

Conflict of interest

None.

Description of authors' roles

Q. Truong designed the study, conducted statistical analyses, and wrote the manuscript. C. Choo supervised the study and edited the manuscript. K. Numbers collaborated with data collection and study design and edited the manuscript. A. Merkin and V. Feigin collaborated with writing and editing the manuscript. N. Kochan assisted with designing instruments, collecting data, and editing the manuscript. P. Sachdev and H. Brodaty sourced funding, supervised data collection, and collaborated with writing and editing the manuscript. O. Medvedev supervised the study and data analyses and edited the manuscript.

Acknowledgments

The Sydney Memory and Ageing Study has been funded by three National Health & Medical Research Council (NHMRC) Program Grants (ID No. ID350833, ID568969, and APP1093083). We thank the participants and their informants for their time and generosity in contributing to this research. We also acknowledge the MAS research team: https://cheba.unsw.edu.au/research-projects/sydney-memory-and-aging-study.

Supplementary material

To view supplementary material for this article, please visit https://doi.org/10.1017/S1041610221000363

References

Allen, M. J. and Yen, W. M. (2001). *Introduction to measurement theory*. Long Grove, IL: Waveland Press.

American Psychology Association (1994). *Diagnostic and Statistical Manual of Mental Disorders*. Washington, DC: American Psychiatric Association.

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

Bloch, R. and Norman, G. (2012). Generalizability theory for the perplexed: a practical introduction and guide: AMEE Guide No. 68. *Medical Teacher*, 34(11), 960–992.

Brennan, R. L. (2010). Generalizability theory and classical test theory. *Applied Measurement in Education*, 24(1), 1–21.

Brodaty, H. *et al.* (2002). The GPCOG: a new screening test for dementia designed for general practice. *Journal of the American Geriatrics Society*, 50(3), 530–534.

Buckley, R. *et al.* (2013). Factors affecting subjective memory complaints in the AIBL aging study: biomarkers, memory, affect, and age. *International Psychogeriatrics*, 25(8), 1307–1315.

Burmester, B., Leathem, J. and Merrick, P. (2016). Subjective cognitive complaints and objective cognitive function in aging: a systematic review and meta-analysis of recent cross-sectional findings. *Neuropsychology Review*, 26(4), 376–393.

Cardinet, J., Johnson, S. and Pini, G. (2011). *Applying Generalizability Theory Using EduG*. New York: Taylor & Francis.

Caselli, R. J. *et al.* (2014). Subjective cognitive decline: self and informant comparisons. *Alzheimer's & Dementia*, 10(1), 93–98.

Chin, Y., Choo, C. C. and Doshi, K. (2019). A case of subjective cognitive complaints in older adults: anxiety, stress, and aging in an elderly client. In: *Clinical Psychology Casebook Across the Lifespan* (pp. 71–77). Singapore: Springer Nature.

Choo, C. C., Harris, K. M., Chew, P. K. and Ho, R. C. (2017). Does ethnicity matter in risk and protective factors for suicide attempts and suicide lethality? *PLoS One*, 12(4), e0175752.

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

Crook, T. H., Feher, E. P. and Larrabee, G. J. J. I. P. (1992). Assessment of memory complaint in age-associated memory impairment: the MAC-Q. *International Psychogeriatrics*, 4(2), 165–176.

Harrison, J. K., Fearon, P., Noel-Storr, A. H., McShane, R., Stott, D. J. and Quinn, T. J. (2015). Informant Questionnaire on Cognitive Decline in the Elderly (IQCODE) for the diagnosis of dementia within a secondary care setting. *Cochrane Database of Systematic Reviews* (3).

Harrison, J. K., Stott, D. J., McShane, R., Noel-Storr, A. H., Swann-Price, R. S. and Quinn, T. J. (2016). Informant Questionnaire on Cognitive Decline in the Elderly (IQCODE) for the early diagnosis of dementia across a variety of healthcare settings. *Cochrane Database of Systematic Reviews*, (11), 1–51.

Hildreth, K. L. and Church, S. J. M. C. (2015). Evaluation and management of the elderly patient presenting with cognitive complaints. *Medical Clinics of North America*, 99(2), 311–335.

Huisman, M. (2000). Imputation of missing item responses: some simple techniques. *Quality and Quantity*, 34(4), 331–351.

Jansen, A. P. D. *et al.* (2008). Self-reports on the IQCODE in older adults: a psychometric evaluation. *Journal of Geriatric Psychiatry and Neurology*, 21(2):83–92.

Jonker, C., Geerlings, M. I. and Schmand, B. (2000). Are memory complaints predictive for dementia? A review of clinical and population-based studies. *International Journal of Geriatric Psychiatry*, 15(11), 983–991.

Jorm, A. (1994). A short form of the Informant Questionnaire on Cognitive Decline in the Elderly (IQCODE): development and cross-validation. *Psychological Medicine*, 24(1), 145–153.

Jorm, A., Scott, R., Cullen, J. and MacKinnon, A. J. P. M. (1991). Performance of the Informant Questionnaire on Cognitive Decline in the Elderly (IQCODE) as a screening test for dementia. *Psychological Medicine*, 21(3), 785–790.

Jorm, A. F. (2004). The Informant Questionnaire on cognitive decline in the elderly (IQCODE): a review. *International Psychogeriatrics*, 16(3), 275.

Langa, K. M. and Levine, D. A. (2014). The diagnosis and management of mild cognitive impairment: a clinical review. *The Journal of the American Medical Association*, 312(23), 2551–2561.

Lenehan, M. E., Klekociuk, S. Z. and Summers, M. J. (2012). Absence of a relationship between subjective memory complaint and objective memory impairment in mild cognitive impairment (MCI): is it time to abandon subjective memory complaint as an MCI diagnostic criterion? *International Psychogeriatrics*, 24(9), 1505–1514.

Medvedev, O. N. *et al.* (2020). A novel way to quantify schizophrenia symptoms in clinical trials. *European Journal of Clinical Investigation*, 51(3), e13398.

Medvedev, O. N., Krägeloh, C. U., Narayanan, A. and Siegert, R. J. (2017). Measuring mindfulness: applying generalizability theory to distinguish between state and trait. *Mindfulness*, 8(4), 1036–1046.

Mitchell, A. J., Beaumont, H., Ferguson, D., Yadegarfar, M. and Stubbs, B. (2014). Risk of dementia and mild cognitive impairment in older people with subjective memory complaints: meta-analysis. *Acta Psychiatrica Scandinavica*, 130(6), 439–451.

Numbers, K., Crawford, J. D., Kochan, N. A., Draper, B., Sachdev, P. S. and Brodaty, H. (2020). Participant and informant memory-specific cognitive complaints predict future decline and incident dementia: findings from the Sydney Memory and Ageing Study. *PLoS One*, 15(5). e0232961.

Park, M. H. (2017). Informant questionnaire on cognitive decline in the elderly (IQCODE) for classifying cognitive dysfunction as cognitively normal, mild cognitive impairment, and dementia. *International Psychogeriatrics*, 29(9), 1461–1467.

Paterson, J. *et al.* (2018). Distinguishing transient versus enduring aspects of depression in New Zealand Pacific

https://doi.org/10.1017/S1041610221000363 Published online by Cambridge University Press

Island children using Generalizability Theory. *Journal of Affective Disorders*, 227, 698–704.

Perroco, T. R. *et al.* (2008). Short IQCODE as a screening tool for MCI and dementia: preliminary results. *Dementia & Neuropsychologia*, 2(4), 300–304.

Petersen, R. C. (2016). Mild cognitive impairment. *CONTINUUM: lifelong Learning in Neurology*, 22(2 Dementia), 404.

Phung, T. K. T. *et al.* (2015). Performance of the 16-item informant questionnaire on cognitive decline for the elderly (IQCODE) in an Arabic-speaking older population. *Dementia and Geriatric Cognitive Disorders*, 40(5–6), 276–289.

Ponds, R. W. and Jolles, J. (1996). Memory complaints in elderly people: the role of memory abilities, metamemory, depression, and personality. *Educational Gerontology: An International Quarterly*, 22(4), 341–357.

Reid, L. M. and MacLullich, A. M. J. (2006) Subjective memory complaints and cognitive impairment in older people. *Dementia and Geriatric Cognitive Disorders*, 22(5–6), 471–485.

Reid, M., Parkinson, L., Gibson, R., Schofield, P., D'Este, C., Attia, J., … and Byles, J. (2012). Memory complaint questionnaire performed poorly as screening tool: validation against psychometric tests and affective measures. *Journal of Clinical Epidemiology*, 65(2), 199–205.

Sachdev, P. S. *et al.* (2010). The Sydney Memory and Ageing Study (MAS): methodology and baseline medical and neuropsychiatric characteristics of an elderly epidemiological non-demented cohort of Australians aged 70-90 years. *International Psychogeriatrics*, 22(8), 1248.

Shavelson, R. J. and Webb, N. M. (1991). *Generalizability Theory: A Primer* (Vol. 1). London, UK: Sage.

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

Slavin, M. J. *et al.* (2015). Predicting cognitive, functional, and diagnostic change over 4 years using baseline subjective cognitive complaints in the Sydney Memory and Ageing Study. *The American Journal of Geriatric Psychiatry*, 23(9), 906–914.

Swiss Society for Research in Education Working Group. (2006). *EDUG User Guide*. Euchatel, Switzerland: IRDP.

Tang, W. K. *et al.* (2004). The scoring scheme of the informant questionnaire on cognitive decline in the elderly needs revision: results of Rasch analysis. *Dementia and Geriatric Cognitive Disorders*, 18(3–4), 250–256.

Truong, Q. C., Krageloh, C. U., Siegert, R. J., Landon, J. and Medvedev, O. N. (2020). Applying generalizability theory to differentiate between trait and state in the Five Facet Mindfulness Questionnaire (FFMQ). *Mindfulness*, 11(4), 953–963. doi: 10.1007/s12671-020-01324-7

Vispoel, W. P., Morris, C. A. and Kilinc, M. (2018). Applications of generalizability theory and their relations to classical test theory and structural equation modeling. *Psychological Methods*, 23(1), 1.

Winblad, B. *et al.* (2004). Mild cognitive impairment–beyond controversies, towards a consensus: report of the International Working Group on Mild Cognitive Impairment. *Journal of Internal Medicine*, 256(3), 240–246.

https://doi.org/10.1017/S1041610221000363 Published online by Cambridge University Press