Journal of Psychology in Africa

ISSN: 1433-0237 (Print) 1815-5626 (Online) Journal homepage: http://www.tandfonline.com/loi/rpia20

Exploring the validity of scores from the Rosenberg Self-Esteem Scale (RSES) in Burundi: A multi-strategy approach

Anne Fromont, Slim Haddad, Rolf Heinmüller, Bruno T. Dujardin & Annalisa Casini

To cite this article: Anne Fromont, Slim Haddad, Rolf Heinmüller, Bruno T. Dujardin & Annalisa Casini (2017) Exploring the validity of scores from the Rosenberg Self-Esteem Scale (RSES) in Burundi: A multi-strategy approach, Journal of Psychology in Africa, 27:4, 316-324

To link to this article: http://dx.doi.org/10.1080/14330237.2017.1347751

Published online: 29 Aug 2017.

Journal of Psychology in Africa, 2017
Vol. 27, No. 4, 316–324, http://dx.doi.org/10.1080/14330237.2017.1347751

Routledge
Taylor & Francis Group

Exploring the validity of scores from the Rosenberg Self-Esteem Scale (RSES) in Burundi: A multi-strategy approach

Anne Fromont[1,*], Slim Haddad[2], Rolf Heinmüller[3], Bruno T. Dujardin[1] and Annalisa Casini[4]

[1]*School of Public Health, Université Libre de Bruxelles, Brussels, Belgium*
[2]*Research Unit, CHU de Québec, Québec, Canada*
[3]*External consultant, Montréal, Québec, Canada*
[4]*Psychological Sciences Research Institute, Université Catholique de Louvain, Louvain-la-Neuve, Belgium*
**Corresponding author email: anne.fromont@ulb.ac.be*

The present study aimed to validate the Rosenberg Self-Esteem Scale (RSES) in Burundi, through a multi-strategy approach used in cross-cultural studies. Respondents were 906 health workers (men = 56%; caregivers 60%). They responded to a bilingual version of RSES. We utilised Confirmatory Factorial Analysis (CFA) with structural equation modelling and a back translation test to explore the structure of the RSES and the reliability of scores from the scale. Data from an independent sample were analysed for the reliability of scores assessment. CFA results suggested that the global RSES factor was likely contaminated by a method-effect; mainly associated with negatively worded items. Internal consistencies and a back-translation test demonstrated that the negatively worded items were unsuitable in this context. The independent sample study confirmed poor reliability and internal consistency of scores for both alternative language versions of the RSES. Our data suggested that an overall cultural effect, rather than a merely specific language effect, may undermine the cross-cultural transportability of the Western scale.

Keywords: Burundi, cross-cultural equivalence, negatively worded item, RSES, self-esteem, structural equation modelling

Introduction

The Rosenberg Self-Esteem Scale (RSES) (Rosenberg, 1965) is one of the most widely used instruments in psychology (Marsh, Scalas, & Nagengast, 2010). It assesses Global Self-Esteem (GSE), or one's overall sense of worthiness as a person. RSES was originally developed in the 1960's to be used with high school students in the United States of America. It has rapidly become the accepted universal gold standard for measuring GSE. Yet, as an important dimension of identity, GSE cannot be considered outside of the specific cultural context since "cultural models of the self define the basis of self-esteem" (Leary & Tangney, 2012, p. 598).

Several methodological and cultural issues raise questions about the cross-cultural transportability of the RSES to African settings. The aim of the present study was to explore the construct validity of RSES for use in Burundi, as well as to determine the reliability of scores from the scale in that country setting. Few studies have examined the validity and reliability of scores from the RSES in African country settings (e.g. Tanzania, DRC and South Africa (Schmitt & Allik, 2005); South Africa (Makhubela & Mashegoane, 2017)). In these studies, the cross-cultural utility of RSES was questioned. In particular, the use of negatively worded items appeared to be problematic in the non-Western culture settings.

Self-construal in Burundi culture

Burundi, Tanzania, and South Africa are predominantly collectivist[1] cultures in which people define and evaluate themselves in primarily interpersonal terms (Haken, Imbriano, Ben Nun, & Tobias, 2011; Heine, Lehman, Markus, & Kitayama, 1999). In such collectivist cultures, shared norms and values prescribe what is socially appropriate. This is inconsistent with the self-definition in individualist cultures, such as in Western countries, in which one's internal attributions are emphasised in seeking information that leads to judgments (Suh, Diener, Oishi, & Triandis, 1998). The sociometer model (Leary, Tambor, Terdal, & Downs, 1995) suggests that self-esteem reflects perceived social acceptance. In such a collectivist culture, social meta-perceptions impact self-appraisal more significantly, and probably differently, than in individualist cultures. These culture-oriented specificities raise questions about the validity of the GSE construct in a collectivist context as it is doubtful that "*one's overall sense of worthiness as a person*" (Schmitt & Allik, 2005, p. 1) is the same construct in both types of cultures or, at least, that it could be assessed in the same way. Nonetheless, to our knowledge, no study using RSES has been conducted in Burundi. The use of RSES in this context raises issues about the validity and reliability of the scale in actually measuring what it purports to measure.

Not all collectivistic cultures are comparable, in particular because of the differences in language nuancing effect on social constructions. Indeed, the perception of expressions such as "*I am a person of worth*" or "*I am a failure*" depends greatly on the semantic paradigm to self-attribute social outcomes. For instance, the concept of self-esteem does not exist in Kirundi language. Instead, personal worthiness is expressed through relational words like "*ubuntu*", which can be translated as "human-ness" or "fraternity"; or "*akarangamutima*", which can be translated as "human value" or "ethics." Thus, in accordance with a collectivistic perception, relational self-identity transcends individual issues and involves the perception of humanist characteristics. Moreover, the specific semantic nuancing calls into question the content validity (i.e., the internal

validity) of indicators of self or person worth premised language variations.

Of course, collectivism and language issues are highly interconnected in an overall cultural paradigm. From a cross-cultural perspective, these could be seen as an etic-emic bind. On the one hand, the use of RSES, regardless of cultural context, implies an etic analysis of a GSE concept considered universally true. On the other hand, self-esteem is, in itself, closely linked to identity and, as a psychological construct, involves individuals *within* specific social groups. Historically, cross-cultural psychological research adopt an emic (within cultures) approach to understanding social phenomenon such as self-esteem (Watkins, 2016; Davidson & Al., 1976). However, much remains for study regarding emic aspects of self-representation.

The RSES: Cross-cultural studies

The RSES was originally designed as a unidimensional self-reporting inventory to measure GSE, but there is controversy regarding its factorial structure. Indeed, a number of scholars have pointed out that the Rosenberg scale contains both positively (P*) and negatively (N*) worded items which could artificially transform GSE into a multidimensional construct. This dichotomy between P* and N* was originally developed to avoid acquiescence bias and relies on the assumption that both kinds of items assess the same construct. However, exploratory factor analysis and Confirmatory Factor Analysis (CFA) have often revealed the presence of two separate factors – associated respectively with P* and N* (Carmines & Zeller, 1979; Corwyn, 2000;Marsh, 1996; Tomàs & Oliver, 1999; Wang, Siegal, Falck, & Carlson, 2001). Some researchers (Ang, Neubronner, Oh, & Leong, 2006;Owens, 1994; Supple, Su, Plankett, Peterson, & Bush, 2012) interpret these factors as two substantially meaningful and distinct dimensions of the RSES, modelled as two separate but correlated latent factors (LFs): negative and positive self-esteem. Using the multitrait-multimethod conceptual framework, other studies support the unidimensionality of the RSES but assert a contamination by a method effect – due to the difference in items' wording. Some of them consider this method effect through the addition, to the unidimensional RSES, of two latent effect factors in a correlated trait – correlated method (CTCM) model.

The CTCM model

The CTCM model is sometimes perceived as a full model (two LFs for the two types of wording) (Marsh, Scalas, & Nagengast, 2010; McKay, Boduszek, & Harvey, 2014; Hyland, Boduszek, Dhingra, Shevlin, & Egan, 2014); and sometimes as a partial and nested model: one LF for the positively or for negatively worded items (DiStefano & Motl, 2006; Horan, DiStefano, & Motl, 2003; Sharratt, Bodusek, Jones, & Gallagher, 2015); and more rarely as a full model where the LFs are correlated (Corwyn, 2000). Yet CTCM models often fail to produce convergent and admissible factor solutions. Thus, correlated trait-correlated uniqueness (CTCU) models are often used (Wang, Siegal, Falck, & Carlson, 2001; Gana, Alaphilippe, & Bailly, 2005; Mullen, Gothe, & McAuley, 2013; Quilty, Oakman, & Risko, 2006; Supple, Su, Plankett, Peterson, & Bush, 2012). The process of correlating uniqueness among positively or negatively worded items, or those worded both ways, certainly leads to difficulties in interpretation. Nonetheless, the CTCM model can handle both unidimensional and multidimensional method effects and at times could be the method of choice (Tomàs & Oliver, 1999).

CFA will allow us to contribute to this debate about the unidimensionality versus multidimensionality of the GSE construct by exploring how different models fit a dataset of Burundian health workers. However, broadly speaking, the many different models reported by the cross-cultural literature on RSES certainly highlight the difficulties of using RSES as a cross-group equivalent scale.

In summary, despite its widespread use, several issues raise serious doubts about the relevance of considering RSES as a universal measure of GSE. The ongoing debate about its factorial structure is consistent with questions of both its construct validity and its content validity across different cultural groups.

Goals of the study

The present study aimed to provide further evidence on the structure of the RSES and reliability of scores from the scale in Burundian culture. We utilised CFA conducted with structural equation modelling (SEM) and internal consistency reliability of scores analysis with a back-translation of test versions. Additionally, we tested reliability and internal consistency of scores from the RSES with an independent sample. This multi-strategy approach prevails in cross-cultural studies (Hui & Triandis, 1985) and can gauge the extent of the equivalence problem of RSES validity in the context of Burundi.

Method

Participants and setting

For the present study, we use data from a longitudinal survey that aims to follow up the impact of an intervention on health services, both on health workers and on the general population in Burundi (Fromont, Heinmuller, & Haddad, 2016). We included all personnel working in 68 health centres (out of 85) in 5 health districts. These health centres volunteered to participate in a German Gesellschaft für Internationale Zusammenarbeit (GIZ) project that aimed to improve the quality of care. A total of 906 health workers (97%) were respondents (mean age = 38.5, SD = 10.27; men = 56% (IC95%: 53-60); caregivers 60% (IC95%: 57-64) and the remaining 40% being support staff). This sample was heterogeneous, as it included various profiles of adults in terms of age, gender, professional function and experience, level of education, and spontaneous language used to respond.

Measures

The RSES is a four point Likert scale, comprised of 10 items; five P* and five N* items[2]. Traditionally, N* are coded in reverse order and all responses are aggregated into a global score from 10 (low GSE) to 40 (high GSE). For the present study, we used the validated French version of the RSES (Vallieres & Vallerand, 1990) to translate

into Kirundi. In order to do so, double-blinded translations were concurrently developed from the French version. The discrepancies between translations were then discussed between the two translators (one of them is a psychologist), another Kirundophone (a Burundian doctor), and the main researcher who also knows the English version well. A back translation from the first Kirundi version helped us to identify problematic issues. At each step, words and expressions were discussed in order to choose the most appropriate translation. Finally, five pre-tests, with health workers in other health districts, provided the opportunity to improve and validate the final version.

Back-translation test

In order to validate cross-cultural equivalence of items in both languages, we independently performed a back-translation test of the final version. It consisted of translating the Kirundi version back to French followed by comparing the original French version and the back-translation result. Discrepancies revealed non-equivalences between items.

The back-translation test confirmed translators' and researchers' preliminary discussions on several difficulties in translating concepts that do not exist in Kirundi. It is noteworthy that only five of the ten items passed the back-translation test; suggesting a different underlying idea and meaning for the five other items. P* items were more successfully translated, as only one of them (P7) showed a discrepancy between the original item "*satisfait de moi*" (i.e. *"satisfied with myself*") and the back-translation item "*fier de moi*" (i.e. "*proud of myself*"). On the contrary, N* items resisted back-translation as four of them revealed a discrepancy between the two versions. Only N5, "*not have much to be proud of"* was successfully translated back. Among the problematic N* items, two raised discussions among translators, and effectively display the cultural issue underlying the wording. Item N3, "*je suis un raté,*" was impossible to translate, as the expression, "*I am a failure*" as a person, does not make any sense for a Burundian. After the back-translation test it became "*je ne vois pas la raison de mon existence*" (i.e. "*I do not see the reason for my existence"*). The same difficulties were encountered with item N8 as "*respect pour moi-même*" (respect for myself) became "*me donner l'honneur*" ("give myself honour", "have honour"). The transition from an individual concept (respect for myself) to a social one (honour) is quite revealing of the collectivist perspective underlying the Kirundi language.

Language preference assessment

We collected qualitative data on participant's language of response preferences, prioritising spontaneous language used to reply. About 31% of the respondents, mainly caregivers (nurses and care assistants) responded to the survey in French, while 69% responded in Kirundi. We note that in analyses of spontaneous language used by groups, professional groups or genders did not show any difference, either for internal consistencies of items or for model results.

Procedure

Permission for the study was granted by the National Committee of Health Research Ethics of Burundi. Participants individually consented to the study. Individual questionnaires were distributed and explained during one of their regular district meetings. As Burundi is a bilingual country, each question was formulated in French and was immediately followed by its Kirundi translation. In order to guarantee anonymity, the return was by individual sealed envelopes and the consent forms were collected separately. The participation rate was 97% with 906 returns recorded from 936 health workers. About 10% of participants indicated that they received help in filling in the questionnaire. These were primarily support staff, probably not very literate, and mainly helped by colleagues. In addition, to an independent sample ($n = 16$), we administered four tests at intervals of a few days and alternated French and Kirundi versions where we mixed the order of the questions.

Analyses: SEM and tests for Goodness of Fit (GoF)

We utilised CFA to test measurement models a priori. The critical issue of these analyses is to determine how well the models fit the data. Several absolute and incremental fit indexes are traditionally used to assess and compare models. We primarily utilised the Root Mean-Square Error of Approximation (RMSEA) to assess for model fit, as well as the Akaike Information Criterion (AIC) to compare alternative models. In addition, we applied the Tucker-Lewis Index (TLI) and the Comparative Fit Index (CFI) to assess for incremental fit indexes that consider how much better the model fits the data compared to a baseline model where all variables are uncorrelated (Mueller & Hancock, 2008). TLI and CFI have to be above 0.95 to accept a model. Finally, we computed the overall coefficient of determination (CD) with an R^2 accounted for variance estimate for the entire model. Models were evaluated using structural equation modelling with Stata 13.1. About 17% of participants had missing data on at least one item, and most of the missing elements were on negatively worded items, especially N3 and N5.

Models were traditionally estimated using Maximum Likelihood (ML), assuming normally distributed multivariate variables. Mardia's coefficient of multivariate kurtosis was 203 and skewness was 54. This indicates a significant non-normality. However, it is recognised that significant non-normality may not impact the results (Mueller & Hancock, 2008). To address problematic multivariate non-normality some authors inspect univariate distribution index values (Mueller & Hancock, 2008; Quilty, Oakman, & Risko, 2006). For our data set, univariate skewness and kurtosis were less than 1.0 and provide evidence that the incidence of non-normality was acceptable. However, both the large sample and the significant non-normality of the multivariate distribution forced us to consider that p-values too close to the classic significance level ($p < 0.05$) were highly suspicious.

Results

Internal consistency of RSES scores

Preliminary exploration of the correlation matrix showed aberrant results for item N8 ("*I wish I could have more respect for myself*") with negative correlations on all other items. As many (7.1%) did not answer this question, it suggests some confusion among them as to its meaning. Since the models including this item gave no meaningful solutions and the back-translation test shows a discrepancy between the underlying concepts in the French and Kirundi version, we decided to eliminate it for the analysis.

Concerning global RSES scale, the reliability of scores was similar with N8 and without N8 (both $\alpha = 0.63$). Reliability of scores was higher among the five P* items ($\alpha = 0.72$) and did not improve after removing the P7 ("*On the whole, I am satisfied with myself*") ($\alpha = 0.70$). Reliability of scores among the four N* was poorest ($\alpha = 0.52$) and did not improve when adding or removing any N* item ($\alpha = 0.38$ to 0.51).

Structure of the RSES: Selected correlated versus uncorrelated models

We identified eight relevant models following SEM (see Figure 1): six *a priori* measurement models from the literature and two final models that improve the fit. Model 1 posits a unidimensional GSE factor (RSES). Model 2 presumes two oblique factors; positive and negative self-esteem as two distinct dimensions. Model 3 treats method effects as two LFs (positive and negative) in addition to a GSE factor (RSES) (CTCM model). Model 4 presumes that a method effect, treated like an LF, affects only N* items. Models 5 and 6 assume partial CTCU respectively on positive and negative items. In addition to these *a priori* measurement models, we explored two models derived from the CTCU model. Model 7 modifies Model 5 by adding a correlation between the uniqueness of items N3 ("*All in all, I am inclined to feel that I am a failure*") and N5 ("*I feel I do not have much to be proud of*") (Lagrange multiplier: 67.5). Model 8 modifies Model 6 by adding a uniqueness correlation between items P2-P7 (Lagrange multiplier: 17.0), P4-P7 (Lagrange multiplier: 14.3), and P4-P6 (Lagrange multiplier: 8.25).

For greater clarity, we have chosen not to present here several other models we explored. We note that the full CTCM model, that considers these two LFs (positive and negative method effects) as correlated, did not provide all fit indexes. This suggests difficulties in reaching a convergent solution. Models that considered method effects such as LFs on positive and negative items respectively (partial CTCM) were with less satisfactory fit than the full CTCU model. Other authors had encountered similar problems (Gana, Alaphilippe, & Bailly, 2005; Wang, Siegal, Falck, & Carlson, 2001) but, even after applying their suggested corrections, these models did not produce a better fit with our dataset.

SEM: Model results

Table 1 reports the fit indexes based on ML for the eight alternative models of the RSES. Consistent with previous research, Model 1 showed the poorest fit with the data, and the addition of any bi-factorial solution always improved the model. This was expected since both the polychoric and usual correlation matrix clearly revealed consistency between P* on the one hand, and N* on the other hand. Models 2, 3, 4, and 5 were also rejected as poor approximations of the data. Model 6, which assumes CTCU between the uniqueness of N* items provides acceptable fit indexes, with the exception of the chi-square which is influenced by both distribution and sample size. This model provided the best fit for our data.

In order to obtain a more satisfactory model, we adjusted Models 5 and 6 by adding a correlation between the forms of uniqueness through the Lagrange multiplier index. This improved all fit indexes for Models 7 and 8, as well as internal reliability. However, several GoF indexes of both models reach caps which make them suspect, leading us to exclude them. It is noteworthy that internal reliability was better with models including method effects as latent factors.

Factor loadings

In further analysis, we examined the factor loading of each item on RSES and the covariance of each element of uniqueness for the best model; i.e. Model 6 (Please see Table 2). For comparison, we also present the results of its LF equivalent (Model 4).

Model 4 showed some strong and significant factor loadings of the P* on RSES. N* items revealed their adverse nature by weak or sometimes non-significant loadings on RSES while the loadings were medium/strong and significant for the negative factors. Standardised variances of uniqueness were high but acceptable (from 0.44 to 0.77), except for N3 and N5 (> 0.80). Model 6, showed the best GoF indexes, and revealed the same structure. However, standardised variances of uniqueness were high but acceptable only for P.*

Reliability and internal consistency of scores from the RSES: Independent sample study

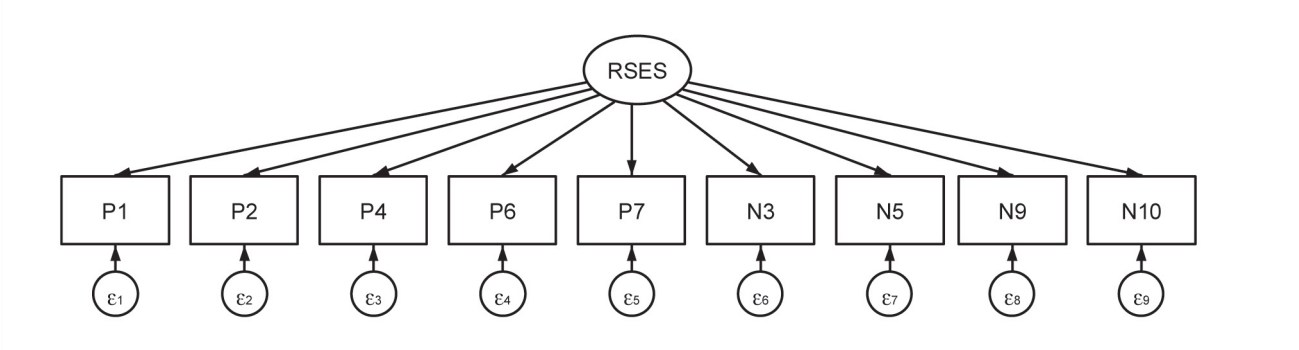
Internal consistency of scores with repeated measures was within acceptable range for the full RSES (Cronbach's alpha about 0.70 to 0.85) and for the N* items (Cronbach's alpha about 0.60 to 0.85) as well as for the P* items (Cronbach's alpha about 0.65 to 0.75).

The French version of the test-retest showed poor correlations for all items, except P7 and N9 (where $r > 0.65$). The Kirundi version of the test-retest also showed poor correlations, although these were generally better than the French version, and only four items (N3, P4, N5, and P7) were acceptable ($r > 0.65$).

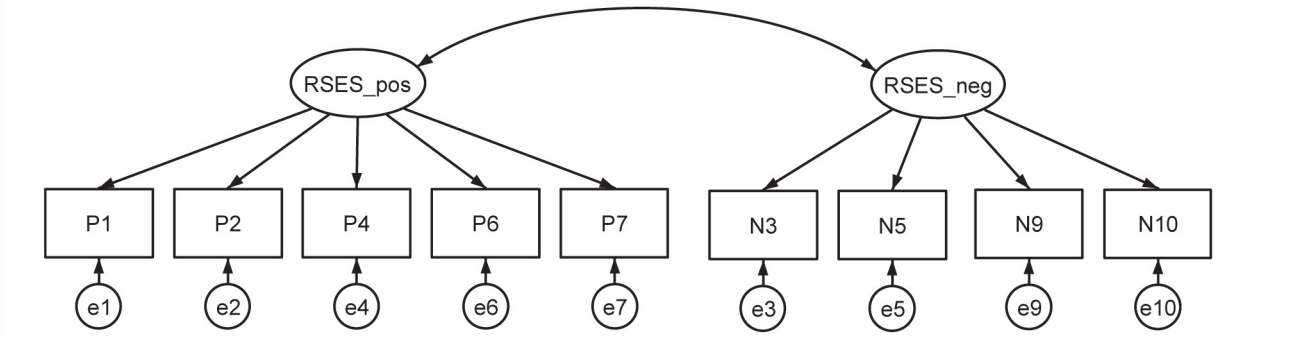
Discussion

The aim of the present study was to explore the cross-cultural validity of the RSES for use with Burundian adults. In our Burundian sample four out of the five original N* showed discrepancies after back-translation. This indicates difficulties grasping these negatively worded concepts in Kirundi. Similar to previous studies, our sample revealed other difficulties with three out of 5 N* items: 10% of N3 and N5 responses were missing while item N8 had to be dropped. According to the RSES literature, respondents typically face a challenge with N* items (DiStefano &

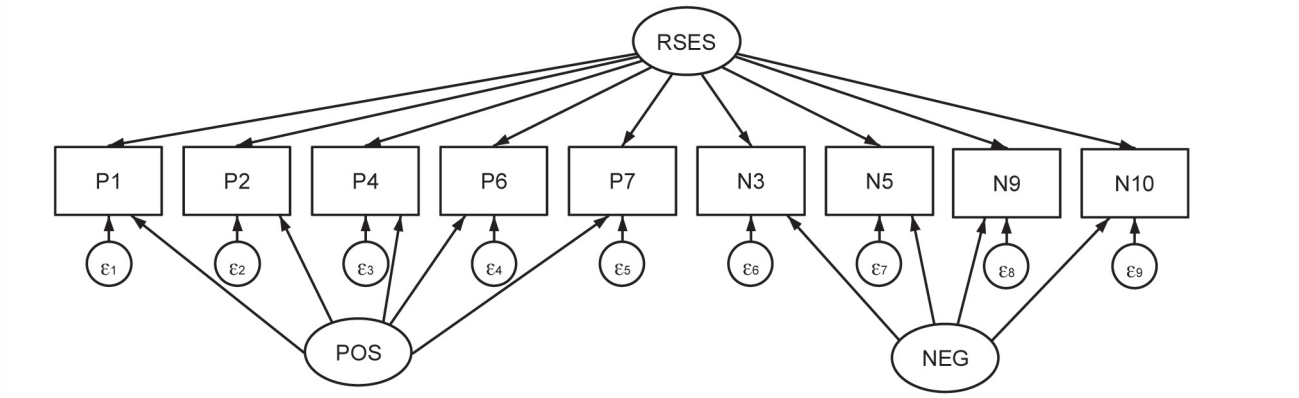
Model 1



Model 2



Model 3



Model 4

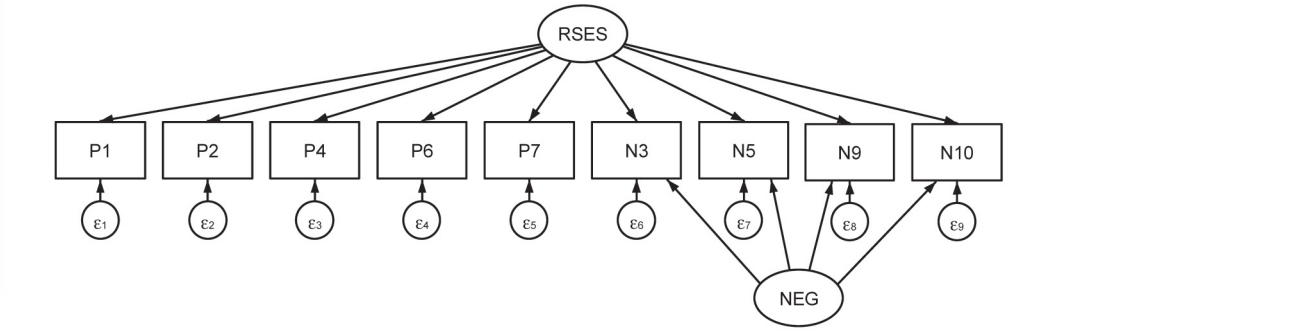
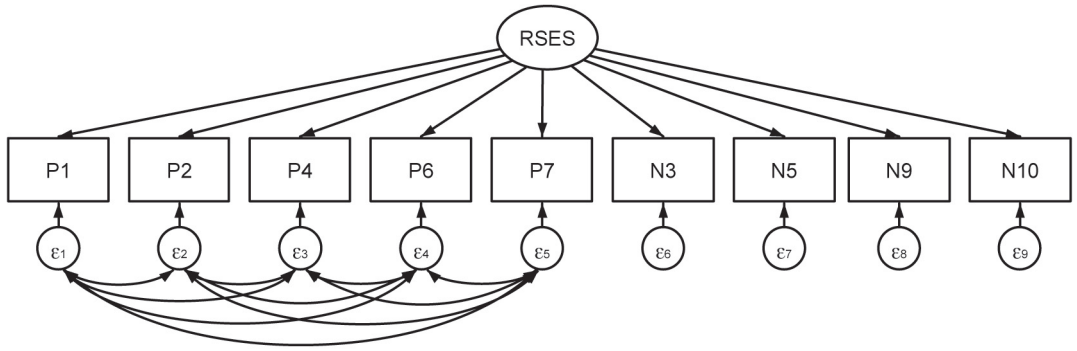
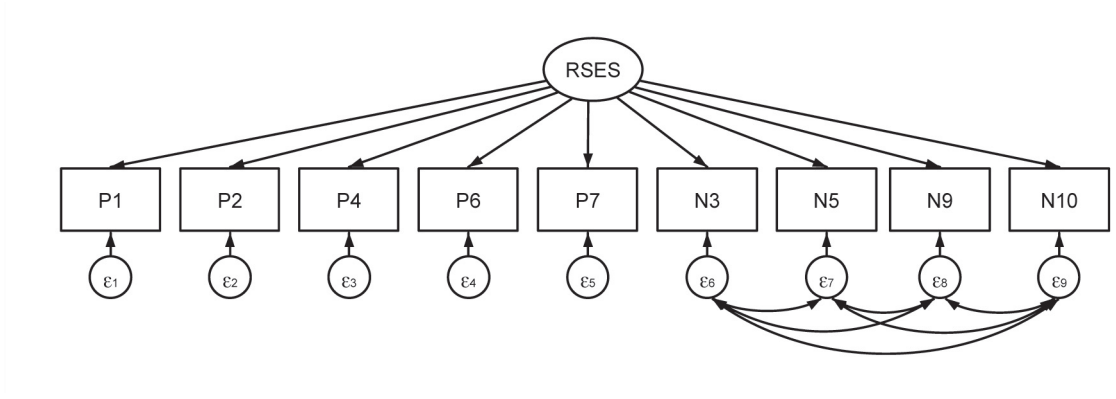


Figure 1. Selected factorial models

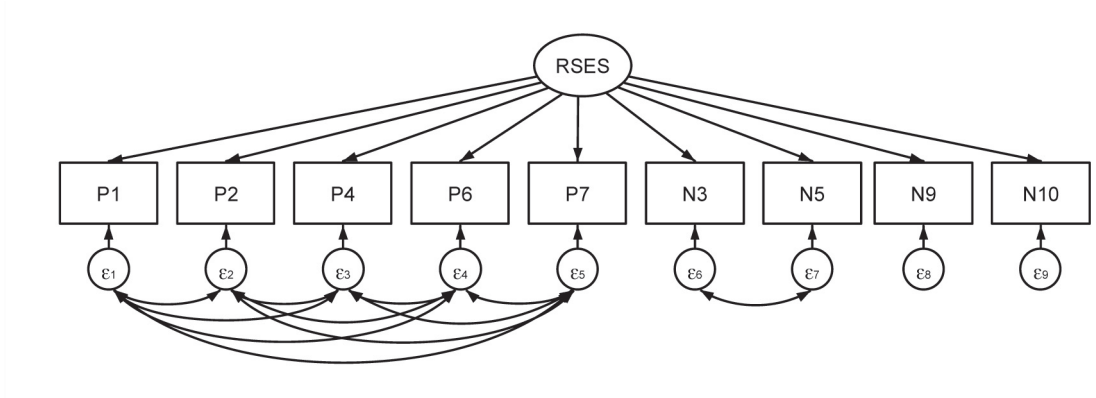
Model 5



Model 6



Model 7



Model 8

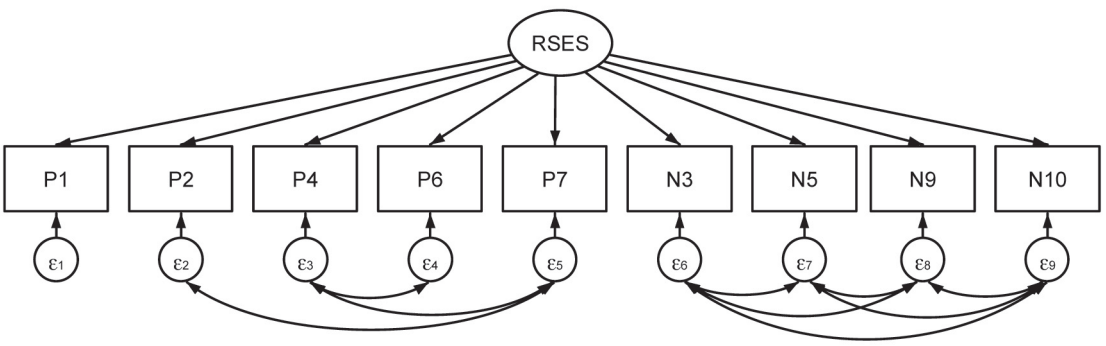


Figure 1. Selected factorial models (cont.)

Motl, 2006; Marsh, Scalas, & Nagengast, 2010; Wang, Siegal, Falck, & Carlson, 2001). Indeed, understanding phrases like "*I wish I could have more respect for myself*" (N8) involves more sophisticated double negative verbal skills than necessary for P*. Double negative logic may be less salient in some language or culture groupings. It also implies cultural context mainly through the semantics of each specific language paradigm. Other researchers with non-western samples have already mentioned this issue for N8 (Ang, Neubronner, Oh, & Leong, 2006; Farruggia, Chen, Greenberger, & Dmitrieva, 2004). In Africa, negative loadings on item N8 were reported in DRC and Tanzania (Schmitt & Allik, 2005). More generally, the use of N* items appears to be a matter of controversy in such cross-cultural settings (Makhubela & Mashegoane, 2017; Schmitt & Allik, 2005). The low reliability of N* items was confirmed in our study by poor and sometimes non-significant loadings on N*. Thus, we assume that negatively worded items are poor indicators of global self-esteem in the Burundi context.

Our results lead us to reject RSES as a full unidimensional scale (Model 1). The correlation matrix and factorial modelling confirmed a clear sub-division between P* and N*. However, the full bi-dimensional model (Model 2) also showed a poor fit with the data. Instead, our results suggested a method-effect on a GSE factor. The last two models retained in our analysis (4 and 6) further suggest that this method-effect is associated with N*, rather than with P*, items. Yet, given the difficulties encountered with the N* items, this method-effect seems to be more related to the overall inadequacy of negative wording, rather than a consistency of meaning. However, given the low reliability of N*, we are not able to assert the validity of scores from the global RSES in the Burundi context. This seems confirmed by the low reliability of both the French and Kirundi version from the test-retest on an external sample.

In sum, with a multi-method approach we were able to test the validity of RSES among an adult population in Burundi. Back-translation tests, internal consistency and CFA revealed that the N* items were unsuitable in this context. The similarity of results regardless, of the spontaneous language used to complete the questionnaire, suggests an overall cultural effect rather than merely a specific language effect.

Conclusion

A multi-method approach allowed us to explore the cross-cultural validity of RSES in the population of adults in Burundi. The N* items proved unsuitable. Indeed,

Table 1. Fit indexes for the eight alternative models of the RSES

Models	X^2	df	X^2/df	*p*	RMSEA	TLI	CFI	AIC	CD
1	353.30	27	13.1	$p < 0.000$	0.115	0.623	0.717	16153	0.756
2	130.08	26	5.0	$p < 0.000$	0.066	0.875	0.910	15932	0.901
3	90.65	18	5.0	$p < 0.000$	0.067	0.874	0.937	15908	0.974
4	128.79	23	5.6	$p < 0.000$	0.071	0.857	0.908	15936	0.899
5	76.43	17	4.49	$p < 0.000$	0.062	0.891	0.949	15896	0.624
6	60.04	21	2.87	$p < 0.000$	0.045	0.942	0.966	15872	0.756
7	7.92	16	0.49	0.951	0.000	1.0	1.0	15830	0.631
8	12.65	17	0.74	0.759	0.000	1.0	1.0	15832	0.824

Table 2. Standardised parameters estimates for models 4 and 6

		Model 4		Model 6	
	Items	Coef.	*p*	Coef.	*p*
Loading on RSES	P1	0.566	$p < 0.000$	0.566	$p < 0.000$
	P2	0.743	$p < 0.000$	0.744	$p < 0.000$
	P4	0.543	$p < 0.000$	0.542	$p < 0.000$
	P6	0.623	$p < 0.000$	0.622	$p < 0.000$
	P7	0.477	$p < 0.000$	0.477	$p < 0.000$
	N3	0.107	0.007	0.114	0.004
	N5	0.052	0.199	0.061	0.130
	N9	0.192	$p < 0.000$	0.191	$p < 0.000$
	N10	0.166	$p < 0.000$	0.1664	$p < 0.000$
Loading on negative factor	N3	0.412	$p < 0.000$	–	–
	N5	0.342	$p < 0.000$	–	–
	N9	0.618	$p < 0.000$	–	–
	N10	0.539	$p < 0.000$	–	–
Covariance between uniquenesses	e.N3*e.N5	–	–	0.358	$p < 0.000$
	e.N3*e.N9	–	–	0.214	$p < 0.000$
	e.N3*e.N10	–	–	0.183	$p < 0.000$
	e.N5*e.N9	–	–	0.165	$p < 0.000$
	e.N5*e.N10	–	–	0.111	0.001
	e.N9*eN10	–	–	0.392	$p < 0.000$

weaknesses appear with these N* items at all stages: back-translation tests, poor internal consistency, and CFA. Given the lack of comparable data from Burundi, we cannot confirm the relevance of RSES nor the reliability of using only P* items. In fact, we can instead confirm that the validity of scores from the RSES remains problematic in the Burundi context. These results underscore the need for meticulous and detailed consideration when transposing concepts and tools in international studies.

In conclusion, our study had to deal with the double challenge of testing the RSES in the context of Burundi and translating this scale into Kirundi. Given the linguistic nature of method effects, our results implied several limitations in terms of both interpretation and replicability.

Authors' note

This research was supported by a scholar grant, for the main author, from the "Fond pour la Formation à la Recherche dans l'Industrie et l'Argiculture" (FRIA) of the Belgian Fund for Scientific Research (FNRS). This publication was supported by the "Fondation Universitaire de Belgique".

End Notes

1 The term "collectivist" refers here to the anthropological concept that has typically divided cultures into two categories: collectivist and individualist (Triandis, 1988).

2 Each item was coded as follows: the letter represents the wording (P for Positively and N for Negatively) and the number represents the order of each item in the French version of the RSES validated by Vallieres & Vallerand (1990).

References

Ang, P., Neubronner, M., Oh, S., & Leong, V. (2006). Dimensionality of Rosenberg's Self-Esteem Scale among normal-technical stream students in Singapore. *Current Psychology (New Brunswick, N.J.), 25*(2), 120–131. https://doi.org/10.1007/s12144-006-1007-3

Carmines, E., & Zeller, R. (1979). *Reliability and Validity Assessment*. Beverly Hills: Sage. https://doi.org/10.4135/9781412985642

Corwyn, R. (2000). The factor structure of global self-esteem among adolescents and adults. *Journal of Research in Personality, 34*, 357-379.

Davidson, A. R., Jaccard, J. J., Triandis, H. C., Morales, M. L., & Diaz-Guerrero, R. (1976). Cross-cultural model testing: Toward a solution of the etic-emic dilemma. *International Journal of Psychology, 11*(1), 1–13. https://doi.org/10.1080/00207597608247343

DiStefano, C., & Motl, R. (2006). Further investigating method effects associated with negatively worded items on self-report surveys. *Structural Equation Modeling, 13*(3), 440–464. https://doi.org/10.1207/s15328007sem1303_6

Farruggia, S., Chen, C., Greenberger, E., & Dmitrieva, J. (2004). Adolescent self-esteem in cross-cultural perspective: Testing measurement equivalence and a mediation model. *Journal of Cross-Cultural Psychology, 35*(6), 719–733. https://doi.org/10.1177/0022022104270114

Fromont, A., Heinmuller, R., Haddad, S. (2016). Accompagnement scientifique et évaluation de l'introduction de mécanismes de management systémique de la qualité dans les services de santé. Final reports from scientific support of the GIZ project *Concours Qualité* (2014–2016).

Gana, K., Alaphilippe, D., & Bailly, N. (2005). Factorial structure of the French version of the Rosenberg Self-Esteem Scale among the elderly. *International Journal of Testing, 5*(2), 169–176. https://doi.org/10.1207/s15327574ijt0502_5

Greenberger, E., Chen, C., Dmitrieva, J., & Farruggia, S. (2003). Item-wording and the dimensionality of the Rosenberg Self-Esteem Scale: Do they matter? *Personality and Individual Differences, 35*(6), 1241–1254. https://doi.org/10.1016/S0191-8869(02)00331-8

Haken, J., Imbriano, J., Ben Nun, A., & Tobias, D. (2011). The Complexities of Culture and Reconciliation in Burundi. *Journal of International Service,* (Spring, Ed.), 35-50.

Heine, S., Lehman, D., Markus, H., & Kitayama, S. (1999). Is there a universal need for positive self-regard? *Psychological Review, 106*(4), 767-794. https://doi.org/10.1037/0033-295X.106.4.766

Horan, P., DiStefano, C., & Motl, R. (2003). Wording effects in self-esteem scales: Methodological artifact or response style? *Structural Equation Modeling, 10*(3), 435–455. https://doi.org/10.1207/S15328007SEM1003_6

Hui, H., & Triandis, H. (1985). Measurement in cross-cultural psychology: A review and comparison of strategies. *Journal of Cross-Cultural Psychology, 16*(2), 131–152. https://doi.org/10.1177/0022002185016002001

Hyland, P., Boduszek, D., Dhingra, K., Shevlin, M., & Egan, A. (2014). A bifactor approach to modelling the Rosenberg Self-Esteem Scale. *Personality and Individual Differences, 66*, 188–192. https://doi.org/10.1016/j.paid.2014.03.034

Leary, M., & Tangney, J. (2012). Culture and Self-Esteem. In *Handbook of Self and Identity* (2nd Ed). New-York: The Guilford Press.

Leary, M., Tambor, E., Terdal, S., & Downs, D. (1995). Self-Esteem as an interpersonal monitor: The sociometer hypothesis. *Journal of Personality and Social Psychology, 68*(3), 518–530. https://doi.org/10.1037/0022-3514.68.3.518

Makhubela, M., & Mashegoane, S. (2017). Psychological validation of the Rosenberg Self-Esteem Scale (RSES) in South Africa: Method effects and dimensionality in black African and white university students. *Journal of Psychology in Africa*. Advance online publication. http://dx.doi.org/10.1080/14330237.2017.1294303

Marsh, H. (1996). Positive and negative global self-esteem: A substantively meaningful distinction or artifactors? *Journal of Personality and Social Psychology, 70*(4), 810–819. https://doi.org/10.1037/0022-3514.70.4.810

Marsh, H., Scalas, F., & Nagengast, B. (2010). Longitudinal tests of competing factor structures for the Rosenberg Self-Esteem Scale: Traits, ephemeral artifacts, and stable response style. *Psychological Assessment, 22*(2), 366-381.

McKay, M., Boduszek, D., & Harvey, S. (2014). The Rosenberg Self-Esteem Scale: A bifactor answer to a two-factor question? *Journal of Personality Assessment, 96*(6), 654–660. https://doi.org/10.1080/00223891.2014.923436

Mueller, R., & Hancock, G. (2008). Best Practices in Structural Equation Modeling. In J. Osborne, *Best Practices in Quantitative Methods* (pp. 488-508). Retrieved from. https://doi.org/10.4135/9781412995627.d38

Mullen, S., Gothe, N., & McAuley, E. (2013). Evaluation of the factor structure of the Rosenberg Self-Esteem Scale in older adults. *Personality and Individual Differences, 54*(2), 153–157. https://doi.org/10.1016/j.paid.2012.08.009

Nunnally, J., & Bernstein, I. (1994). *Psychometric Theory* (3rd ed.). New York: McGraw-Hill.

Owens, T. (1994). Two dimensions of self-esteem: Reciprocal effects of positive self-worth and self-deprecation on adolescent problems. *American Sociological Review, 59*(3), 391–407. https://doi.org/10.2307/2095940

Quilty, L., Oakman, J., & Risko, E. (2006). Correlates of the Rosenberg Self-Esteem Scale method effects. *Structural Equation Modeling, 13*(1), 99–117. https://doi.org/10.1207/s15328007sem1301_5

Rosenberg, M. (1965). *Society and the Adolescent Self-Image.* Princeton: Princeton University Press

Schmitt, D. P., & Allik, J. (2005). Simultaneous administration of the Rosenberg Self-Esteem Scale in 53 nations: Exploring the universal and culture-specific features notions of global self-esteem. *Journal of Personality and Social Psychology, 89*(4), 623–642. https://doi.org/10.1037/0022-3514.89.4.623

Sharratt, K., Bodusek, D., Jones, A., & Gallagher, B. (2015). (Accepted for publication). Construct validity, dimensionality and factorial invariance of the Rosenberg Self-Esteem Scale: A bifactor modelling approach among children of prisoners. *Current Issues in Personality*.

Suh, E., Diener, E., Oishi, S., & Triandis, H. (1998). The shifting basis of life satisfaction judgments across cultures: Emotions versus norms. *Journal of Personality and Social Psychology, 74*(2), 482–493. https://doi.org/10.1037/0022-3514.74.2.482

Supple, A., Su, J., Plankett, S., Peterson, G., & Bush, K. (2012). Factor structure of the Rosenberg Self-Esteem Scale. *Journal of Cross-Cultural Psychology, 44*(5), 748-764.

Tomàs, J., & Oliver, A. (1999). Rosenberg's Self-Esteem Scale: Two factors or method effects. *Structural Equation Modeling, 6*(1), 84–98. https://doi.org/10.1080/10705519909540120

Triandis H. (1988) Collectivism v. Individualism: A Reconceptualisation of a Basic Concept in Cross-cultural Social Psychology (pp 60-95). In Verma G.K., Bagley C. (eds) *Cross-Cultural Studies of Personality, Attitudes and Cognition*. Palgrave Macmillan, London.

Vallieres, E. F., & Vallerand, R. J. (1990). Traduction et validation canadienne-française de l'echelle de l'estime de soi de Rosenberg. *International Journal of Psychology, 25*(2), 305–316. https://doi.org/10.1080/00207599008247865

Wang, J., Siegal, H., Falck, R., & Carlson, R. (2001). Factoral structure of Rosenberg's Self-Esteem Scale among crack-cocaine drug users. *Structural Equation Modeling, 8*(2), 275–286. https://doi.org/10.1207/S15328007SEM0802_6

Watkins, D. (2016). Within-culture and gender differences in self-concept. *Journal of Cross-Cultural Psychology, 26*(6), 692–699.