Voluntas (2016) 27:1403–1424
DOI 10.1007/s11266-015-9602-2

ORIGINAL PAPER

Validating Grant-Making Processes: Construct Validity of the 2013 Senior Corps RSVP Grant Review

Erwin Tan[1] · Robin Ghertner[1] · Patricia J. Stengel[1] · Malcolm Coles[1] · Vielka E. Garibaldi[1]

Published online: 6 June 2015

Abstract Accountability in grant-making requires a valid, fair and transparent selection process. This study proposes a four-step framework for validating such a process: determine standards for qualified applicants, assess inter-reviewer reliability, assess factorial validity, and assess reliability. This framework is applied to the Corporation for National and Community Service's 2013 RSVP grant-making process. The standards were close to the highest points of reliability. Inter-reviewer reliability was above 0.90, a common threshold for high-stakes measurement. After conducting confirmatory factor analysis, the final model merged two of the original five domains of selection criteria, resulting in four domains. The final model was found to have strict measurement invariance, high convergent validity, and measurement reliability between 0.88 and 0.93 for all domains. The results validate the 2013 review process and indicated that the scores exhibited high degrees of reliability, giving public assurance that the process was sufficiently objective and accurately reflected program priorities.

The authors prepared this study as part of their official duties as employees of the Corporation for National and Community Service 1201 New York Avenue, NW, Washington, DC 20525, USA.

✉ Robin Ghertner
robinghertner@gmail.com

Erwin Tan
etan@cns.gov

Patricia J. Stengel
pstengel@cns.gov

Malcolm Coles
mcoles@cns.gov

Vielka E. Garibaldi
vgaribaldi@cns.gov

[1] Corporation for National and Community Service, 1201 New York Avenue, NW, Washington, DC 20525, USA

Résumé La responsabilité en matière d'octroi de subventions nécessite un processus de sélection valide, équitable et transparent. Cette étude propose un cadre en quatre étapes pour la validation de ce processus : déterminer les critères pour les demandeurs qualifiés, évaluer la fiabilité entre les examinateurs, évaluer la validité factorielle et évaluer la fiabilité. Ce cadre est appliqué au processus d'octroi de subventions RSVP de 2013 de la *Corporation for National and Community Service*. Les critères étaient proches des points les plus hauts de fiabilité. La fiabilité entre les examinateurs était supérieure à 0,90, un seuil commun pour la mesure des enjeux majeurs. Après avoir procédé à une analyse factorielle confirmatoire, le modèle final a combiné deux des cinq domaines originaux des critères de sélection, ce qui a conduit à quatre domaines. Il a été constaté que le modèle final avait une invariance de mesure stricte, une validité convergente élevée et une fiabilité de mesure entre 0,88 et 0,93 pour tous les domaines. Les résultats valident le processus d'examen de 2013 et ont indiqué que les points présentaient des degrés élevés de fiabilité, ce qui donne aux citoyens l'assurance que le processus était suffisamment objectif et qu'il reflétait fidèlement les priorités du programme.

Zusammenfassung Die Rechenschaftslegung bei der Vergabe von Fördermitteln erfordert ein gültiges, faires und transparentes Auswahlverfahren. Diese Studie schlägt ein Vier-Stufen-Rahmenwerk zur Validierung eines solchen Verfahrens vor: das Festlegen von Standards für qualifizierte Bewerber, die Bewertung der Zuverlässigkeit interner Prüfer, die Bewertung der faktoriellen Validität und die Bewertung der Zuverlässigkeit. Dieses Rahmenwerk wird auf das 2013 von der RSVP-Organisation der Corporation for National and Community Service durchgeführte Verfahren zur Vergabe von Fördermitteln angewandt. Die Standards erreichten beinahe die höchsten Messwerte für die Zuverlässigkeit. Die Zuverlässigkeit interner Prüfer lag über 0,90, ein üblicher Grenzwert für höchst relevante Messungen. Nach Durchführung einer konfirmatorischen Faktorenanalyse wurden in dem letztendlichen Modell zwei der ursprünglich fünf Bereiche der Auswahlkriterien zu einem Bereich zusammengefasst, so dass am Ende vier Bereiche vorlagen. Das endgültige Modell wies für alle Bereiche eine strikte Messungsinvarianz, eine höchst konvergente Validität und eine Messzuverlässigkeit zwischen 0,88 und 0,93 auf. Die Ergebnisse validieren das Prüfverfahren von 2013 und zeigen, dass die Werte ein hohes Maß an Zuverlässigkeit darstellen, wodurch öffentlich versichert wird, dass das Verfahren ausreichend objektiv war und die Programmprioritäten korrekt widerspiegelte.

Resumen La responsabilidad en la concesión de subvenciones requiere un proceso de selección válido, justo y transparente. El presente estudio propone un marco de cuatro pasos para validar dicho proceso: determinar normas para los demandantes cualificados, evaluar la fiabilidad entre revisores, evaluar la validez factorial, y evaluar la fiabilidad. Este marco se aplica al proceso de concesión de subvenciones RSVP 2013 de la Corporation for National and Community Services (Corporación para Servicios Comunitarios y Nacionales). Las normas estuvieron muy cerca de los puntos más altos de fiabilidad. La fiabilidad entre revisores estuvo por encima de 0,90, un umbral común para la medición de alta exigencia. Después de

realizar el análisis confirmatorio de factores, el modelo final fusionó dos de los cinco campos originales de criterios de selección, dando lugar a cuatro campos. Se encontró que el modelo final tenía una invarianza de medida estricta, una fiabilidad de medición y validez altamente convergentes entre 0,88 y 0,93 para todos los campos. Los resultados validan el proceso de revisión de 2013 e indican que las puntuaciones mostraban altos grados de fiabilidad, ofreciendo la garantía pública de que el proceso fue suficientemente objetivo y reflejó con precisión las prioridades del programa.

Keywords Grant-making · Accountability · Decision-making · Validity · Transparency

Introduction

The increasing pressure to make grant-making accountable has pushed governmental, non-profit, and private sector grant-makers to focus on the outcomes and impact of their funding decisions. However, there has been limited emphasis on the strength of the actual decisions themselves. Accountability in grant-making requires an objective and transparent selection process. Arguably, awarding grants in a manner that accurately and precisely aligns with the funder's objectives is a prerequisite to ensuring that grantee outcomes are aligned in such a way.

The decision to fund an applicant is in essence a question of measurement: How to define a procedure that measures the likelihood that an applicant will be a successful grantee in a valid and reliable way? Psychometricians over the past three decades have developed rigorous methods to assess measurement procedures (McDonald 1999). Some of these tools have found prominence in measuring the inter-reviewer reliability of peer review processes (Bornmann et al. 2010; Kotchen et al. 2004; Marsh et al. 2008; Mutz et al. 2012; Rothwell and Martyn 2000). However, this is only one component of measurement, and a detailed literature search did not find a single study that applies the full range of measurement tools to grant-making. This study uses a four-step process, outlined in Fig. 1, to assess the grant-making decisions in the 2013 RSVP Grant Competition, run by the Corporation for National and Community Service (CNCS). Additional steps to validate the grant review process include content validation and predictive or concurrent validation (Haynes and Kubany 1995; McDonald 1999), and ongoing research is being conducted to assess these types of validity of CNCS grant-making processes.

The first step defines the standards, a priori, for minimally qualified and best-qualified applications that are eligible for funding, and operationalizes them as

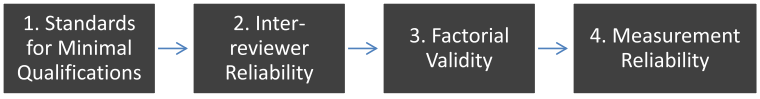

Fig. 1 Four-step framework for establishing construct validity and reliability of a grant-making decision

passing scores. The second step assesses the inter-reviewer reliability between application reviewers. This establishes the congruence between reviewers in their understanding and ratings of applications. The third step assesses the factorial validity of the selection criteria, which evaluates the extent to which they accurately reflect the underlying common constructs or dimensions. The final step assesses the measurement reliability of the grant scores.

Background

CNCS is the federal agency for domestic civilian national service, funding programs such as AmeriCorps, Senior Corps, and the Social Innovation Fund. RSVP is one of the three major Senior Corps programs, and one of the largest volunteer programs targeting senior citizens in the US, offering a diverse range of volunteer activities serving communities across the country. In fiscal year 2013, RSVP engaged over 274,500 volunteers who served in more than 38,000 community organizations, with an annual federal appropriation of over $47 million. RSVP provided independent living services to 610,000 adults, respite services to nearly 15,000 family or informal caregivers and mentored more than 87,000 children (CNCS 2013a, b, c). RSVP volunteers serve with commitments ranging from a few hours to 40 h per week. RSVP grantees receive funding for the recruitment, placement, and coordination of volunteers ages 55 and older in a specified geographic area in which they are the sole RSVP program.

RSVP grants are awarded based on geographic funding areas, and CNCS issues only one RSVP grant in each area. After being initially competitively awarded in 1971, until 2013 RSVP grants were non-competitively renewed. The 2009 Edward M. Kennedy Serve America Act (Serve America Act) authorized CNCS to use a competitive process to award RSVP grants from fiscal years 2013 through 2015. This analysis focuses on the review that took place for the 2013 funding year, involving awards previously granted to 240 incumbent RSVP grantees whose grant cycle would end in 2013 (36 % of the then active RSVP grant portfolio).

Prior to initiating the grant competition, Senior Corps engaged with current grantees to provide feedback on competition planning and issued updated RSVP program regulations. Senior Corps staff provided additional technical assistance and training to all incumbent grantees as required by the Serve America Act, including providing them a customized evaluation in advance of the competition to identify strengths, challenges, and training and technical assistance needs, and also included stakeholder input (Senior Corps 2010). Details of the award process and instructions to applicants were published in the 2013 RSVP Notice of Funding Opportunity (henceforth, 2013 Notice) (CNCS 2012).

Grant Review Process

Grant reviewers rated applications on 23 selection criteria, listed in the "Appendix," reflecting legislative and policy priorities in the following categories: Program Design, Organizational Capacity, and Cost Effectiveness and Budget Adequacy

Table 1 Conceptual model of application quality

Category	Sub-category (domains)	Number of criteria	Percent of total score (%)
Program design	Strengthening communities	6	20
	Recruitment and development of volunteers	4	15
	Performance measure work plan	1	15
Organizational capacity	Program management	5	15
	Organizational capability	5	20
Cost effectiveness and budget adequacy		3	15

(CNCS 2013a, b, c). In developing the criteria, senior staff divided the three categories into 6 sub-categories that represented specific domains (see Table 1). The selection criteria were designed to reflect dimensions of each domain that could (1) be answered in an application (2) discriminate a strong applicant from a weak applicant, and (3) establish a minimum fundable score under which applicants represented an unacceptable level of programmatic and financial risk. All criteria but one were designed to be reviewed by a panel of external and internal reviewers. The remaining criterion, 'National Performance Measure outcome work plans above the minimum 10 %,' was automatically scored by the CNCS performance measurement data system. Reviewers could ask clarifying questions about this criterion but were instructed not to rate it.

To determine which applications would be funded, Senior Corps used a criterion-referenced standard based on a definition of minimally qualified and best qualified, described in more detail below (Cizek and Bunch 2007). Selection criteria and their weights were developed by program leadership in charge of policy and administration, with further input from CNCS leadership and the Office of Management and Budget. The selection criteria were communicated to applicants in application instructions and the 2013 Notice.

The Serve America Act directed CNCS to use a blended review of staff reviewers and peer reviewers 'including members with expertise in senior service and aging, to review applications' for RSVP grants. Reviewers were organized into 43 panels, each panel including 2 internal staff reviewers and 1 external peer reviewer, for a total of 86 staff reviewers and 43 peer reviewers (Senior Corps 2013a). Each panel reviewed between five and six applications; no application was reviewed by more than 1 panel. Staff reviewers only reviewed applications for funding opportunities outside their state to ensure that reviewers were not biased through previous contact with the applicants. CNCS recruited peer reviewers on the basis of

(1) Minimum of 5 years of applicable experience with adults 55 and older or a minimum of 2 years of applicable experience with older adults and a 4-year college degree;
(2) Minimum of 2 years of experience in any of the CNCS Focus Areas;
(3) Good oral and written communication skills; and
(4) The ability to collaborate with peers.

All peer and staff reviewer candidates were screened for any potential conflicts of interest, and received 5 h of training. Training topics included an overview of CNCS and RSVP, preparation for the grant application review, how to review against the selection criteria, how to prepare comments, and finally setting expectations for reviewers. All reviewers were instructed to be familiar with the 2013 Notice, grant application instructions, frequently asked questions document, and RSVP regulations. Reviewers were provided with an RSVP Reviewer Handbook and a sample application exercise and a sample completed individual reviewer form.

The grant review process spanned slightly over three weeks and required approximately 50 h of time from each reviewer to complete. For each application, reviewers first completed their reviews independently. Then all three reviewers held a panel discussion call to discuss their assessment of the application. The purpose of the discussion was not to arrive at consensus, but to ensure that all reviewers understood the application in the same way and understood how to apply the selection criteria. Following the call, reviewers were given the opportunity to update their ratings based on the call discussion, but consensus was not required. The final scores used to judge applicants were created by summing weights for the final ratings from each reviewer, and then taking their average.

Methods

Scoring and Standards for Minimal and Best Qualifications

The first step in our 4-step framework sets standards and passing scores for applications eligible for funding. The method for determining passing scores is essential to establishing the overall validity of the grant-making decision. Valid scores do not translate to a valid decision if the way those scores are used does not reflect the precision and accuracy they represent. Senior Corps designed two standards for the RSVP competition: minimally qualified and best qualified.

Two standards were established: Minimally Qualified and Best Qualified. Minimally Qualified was defined as applications with an acceptable degree of confidence in their success as a grantee, with sufficient quality across all selection criteria. These applications were required to represent reasonable plans but could sometimes be unclear about a specific part of their applications, meaning that the reviewers thought the applications made some assumptions. Best Qualified was defined as applications with a high degree of confidence of success, by meeting or exceeding the standards of most of the criteria. These applications were required to represent reasonable plans that provided all of the required information, meaning that the reviewers thought that the applicant explained most of their assumptions and reasons.

These two passing scores were operationalized in the scoring rubric for the criteria (Senior Corps 2012). The rubric contained four levels: ‘Excellent,’ ‘Good,’ ‘Fair,’ Does not meet.’ To be rated ‘Excellent,’ applicants must go beyond what is requested by the selection criteria. To be rated ‘Good,’ applicants must address

everything requested in the selection criteria. The passing score for Minimally Qualified was set at the score for meeting the equivalent of 'Fair' on all criteria. The passing score for 'Best Qualified' was set at the score for meeting the equivalent of 'Good' on all criteria. Scoring was compensatory, meaning that applicants could make up for low assessment on one criterion by a high assessment on another.

To assess the reliability of these standards, the amount of information provided by reviewers was assessed at the passing scores associated with Minimally Qualified and Best Qualified using a graded response model (GRM). A graded response model is a type of item response model with ordinal indicators, and is functionally equivalent to a confirmatory factor analysis (Muthén and Asparouhov 2005). The GRM, and item response theory in general, allows the calculation of the information content of a rating procedure at different quality levels, where information is defined as the inverse of the variance of factor scores. Unless all selection criteria have the same statistical parameters in the GRM (i.e., discrimination, thresholds between response options), any given observed score can be obtained by multiple response patterns across the criteria. This means that factor scores do not match one to one with the observed passing score. As a result, to estimate the information content at the passing scores, the observed passing score must be converted to a range of factor scores that can obtain the same passing score.

The ideal situation would be that the passing scores used to define minimally qualified and best qualified are at the points of maximal information. This means the rating procedure would be most reliable at the passing scores, providing confidence that the scores are good at differentiating between applications that meet the requirements and those that do not.

Inter-Reviewer Reliability

Inter-reviewer reliability assesses the alignment of reviewer ratings. This alignment has two components: consensus and consistency (Stemler 2014). Consensus means the extent to which reviewers come to exact agreement on an application. For example, a perfectly reliable scale in the consensus aspect would have all reviewers give the exact same rating to applications. Consistency means the degree to which reviewers consistently rate applications. For example, a perfectly reliable scale in the consistency aspect means that all reviewers rate one application high and another low, even if they do not provide the exact same rating to each. Scales that are reliable in the consensus aspect are also reliable in the consistency aspect, but not vice versa. Krippendorf's α (Krippendorff and Bock 2008) was chosen to measure consensus and Cronbach's α (Cronbach 1951) to assess consistency. While Krippendorf's α is not the most common consensus metric, it is the only metric specifically designed to deal with ordinal and continuous data, as well as more than two reviewers. Cronbach's α is commonly used for ordinal, continuous, and dichotomous data.

Reliability is measured on a 0–1 scale, where 1 indicates perfect reliability and a 100 % likelihood of the same scores being produced in a separate rating procedure, while 0 indicates complete lack of reliability. The ratings on individual criteria are treated as medium stakes, and therefore, 0.70 was used as a guide for reliability.

Since the consequences of the overall rating score are that an organization will be funded or not, the ratings for the overall score are considered high stakes and use 0.90 as the threshold for reliability (Krippendorff and Bock 2008; Nunnally and Bernstein 1978).

As described above, reviewers on the same panel held a discussion before submitting their final scores. While the discussion was not intended to lead to consensus decisions, and in most cases did not, it did likely cause reviewers with disparate scores to adjust their scores closer together. Because of this, it is expected to find a high level of inter-reviewer reliability.

Factorial Validity

Factorial validity is established in three steps: dimensionality, invariance, and convergence and discrimination. Dimensionality assesses how many concepts are being measured by each category. Principles of measurement require that categories should be unidimensional, meaning that each category measures a single construct or characteristic (Bond and Fox 2013; McDonald 1999). To assess unidimensionality, an exploratory factor analysis was first conducted to identify whether criteria loaded on the categories as outlined in Table 1. Then a confirmatory factor analysis (CFA) was conducted on the five category model, examining fit indices, factor loadings, and modification indices. CFA is one of the most common methods to assess dimensionality, and is advantageous because its methods, strengths, and limitations are well established (Takane and de Leeuw 1987). The Lavaan package in R was used to estimate the CFA, using means and variance adjusted weighted least squares with robust standard errors (Beauducel and Herzberg 2006; Rosseel 2012). CFA relies on either the covariance or correlation matrix of the variables, and because our data are ordinal it is necessary to use polychoric correlation, which is designed for this data type. After estimating the polychoric correlation matrix of the criteria, the algorithm used that matrix as the inputs for CFA estimation. After estimating the initial model, certain constraints were applied to test for improved model fit or increased parsimony, which would simplify the interpretation of the model. These constraints included merging categories, changing paths between criteria and categories, and constraining loadings across criteria within a category.

The second step to establish factorial validity is to assess measurement invariance—that is, whether the same factor model fits well for different subgroups (Meredith, 1993). This was done across two types of groups. The first was based on the dollar amount of funding opportunities, where two funding groups were created—high funding and low funding—by splitting the funding opportunities at the median. The second grouping was based on the service focus area of grantees. Grantees are asked to select their primary focus area among the six focus areas outlined in the Serve America Act, including Disaster Services, Economic Opportunity, Education, Environmental Stewardship, Healthy Futures, and Veterans and Military Families. Eighty percent of grantees selected Healthy Futures for their primary focus area, so measurement invariance was analyzed between this group and applicants that chose other focus areas. The approach to assess invariance followed the strategy outlined by Meredith (1993) and Vandenberg (2002), which

goes through a series of tests of invariance, from least to most strict (Chandra, 2011). Configural invariance tests that the overall factor structure is the same across groups; weak invariance tests that the loadings are the same; strong invariance tests that the loadings and the thresholds are the same; and strict invariance tests that the loadings, thresholds, and residuals are the same (Muthén and Asparouhov 2005).

The third step to establish factorial validity assesses the convergence and discrimination (Hair et al. 2010). Convergence refers to the degree to which the criteria correlate well with each other within the category they reflect. Fornell and Larcker (1981) and Hair et al. (2010) state that convergence is established if the dimension's reliability is greater than the average variance extracted. The second criterion for convergence is that the average variance extracted is greater than 0.50, meaning more than half of variance of the criteria is explained by their respective category. Discrimination refers to the degree to which each category is better explained by its own criteria than by the criteria from another category—essentially how well the category discriminates from other categories. Hair et al. (2010) further state that discrimination is established if the average variance extracted is greater than the maximum shared variance and the average shared variance of the underlying categories. This means that a category has a stronger relationship with its indicators than with any of the other categories (Fornell and Larcker 1981; Hair et al. 2010).

Measurement Reliability

Measurement reliability means the degree to which an instrument precisely measures what it is intended to measure, and is generally defined as the ratio of true variance to total variance (Lord et al. 1968). There are a number of different measures of reliability designed for different types of factor models. The most common measure—Cronbach's α—requires that the each category in the model is essentially τ-equivalent. Essential τ-equivalence means that the model's items measure the same construct, on the same scale; in a CFA, this means the loadings of each item on the category are equivalent, though their error variances may be different (Graham, 2006). If the model is congeneric, meaning the items measure the same construct but on different scales (i.e., the factor loadings are not the same), Cronbach's α underestimates reliability. In this case, ω provides a better estimate (McDonald 1999; Raykov 2001). In addition, Cronbach's α and most other estimates of reliability require the constructs to be unidimensional, meaning they measure a single concept (Meyer 2010). However, overall RSVP application quality has been designed as a multidimensional construct, and thus, a more robust method is required. Multidimensional ω was used to assess the measurement reliability of the entire instrument, which is designed for multiple dimensions and takes account of the model-explained and error variances across all categories (Fornell and Larcker 1981; Gignac 2014; Graham 2006; Raykov 2001; Revelle and Zinbarg 2008). Measurement reliability was assessed using the 0.90 cutoff value for the entire instrument, and 0.70 for each category.

Table 2 Descriptive summary of sample

Number of applicants	241
Number of reviewers	132
Number of panels	44
Number of criteria	22
Number of total reviewer-application combinations	723
Mean total score (standard deviation), on a 0–88 scale	57 (12.26)

Sample Characteristics

Data for this study came from the ratings of 241 applications by 132 reviewers. Each application was reviewed by three reviewers, and rated on 22 criteria. This resulted in 723 unique reviewer-application combinations and 15,906 unique criterion-reviewer-application combinations. As stated above, a 23rd criterion was not rated by reviewers and therefore was not analyzed. The reviewers were clustered into 44 panels, with each panel reviewing between five and six applications. Table 2 summarizes these characteristics.

External and internal reviewers did not differ systematically in how they rated applications: using a multilevel regression model of the total applicant score, external reviewers on average scored applications about 2 points higher. While this difference is statistically significant, it represents a small difference on an 88 point scale, and is not sufficient to have moved many inadequate applications above the threshold. Descriptive statistics on each criterion can be found in the "Appendix."

Results

Standards for Minimal and Best Qualifications

The overall test information content was near its highest at the point of the two passing scores for minimally qualified and best qualified. Figure 2 shows the information associated with different factor scores as estimated by the graded response model.

There are three clear peaks in the information curve, which could be natural points for setting passing scores. For the first peak, the average rating across all criteria would be just over the lowest rating, 'Does not meet.' For the second peak, the average rating would be just over 'Fair,' and the third peak would be just over 'Good.' This aligns well with the passing scores determined in the grant application review process, which were defined at 'Fair' and 'Good.' The exact range of factor scores associated with the minimal-qualified and best-qualified passing scores are highlighted in dark gray and light gray, respectively. Although neither passing score is at the highest point of test information, setting the average ratings exactly at *fair* and *good* has more face validity than creating a passing score based on complex response patterns to the criteria. This result indicates that the passing scores

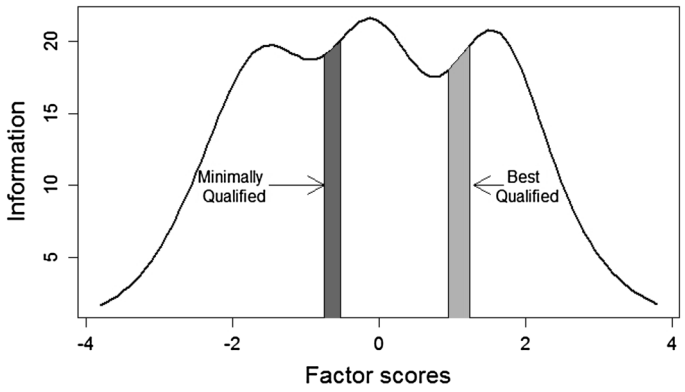

Fig. 2 Test information curve and passing scores

determined in the grant application review process exhibit close to the maximum reliability of the rating procedure.

Inter-Reviewer Reliability

On both consensus and consistency measures, the reliability of the overall score was over 0.90, as shown in Table 3. On average, the reliability of individual criteria was lower than the overall score, although there was considerable variability across them.

Figure 3 provides the distribution for the criteria on both reliability coefficients, which shows the criteria performed worse on the consensus measures. All criteria were above the 0.7 threshold on Cronbach's α, and half were above it for Krippendorf's α.

Factorial Validity

As discussed above, our initial modeling strategy tested for the dimensionality of the factor model, by estimating an initial model based off of Table 1 and then applying constraints to improve the model or simplify interpretation. The fit of these models is reported in Table 4.

The initial factor model resulted in a borderline fit, with the RMSEA at 0.07, statistically different from the commonly applied 0.05 threshold. The CFI and TLI were fairly high, but the modification indices indicated that the model could be

Table 3 Inter-reviewer reliability for reviewer ratings

Reliability coefficient	Overall score	Mean reliability of criteria
Krippendorf's α	0.91	0.71
Cronbach's α	0.97	0.88

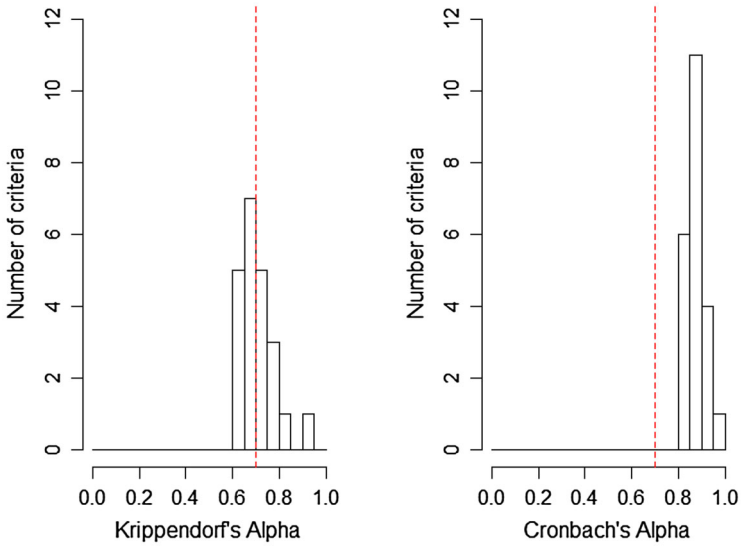

Fig. 3 Distributions of inter-reviewer reliability coefficients for 22 criteria. *Dashed lines* represent the cutoff threshold for adequate reliability, set at 0.7 based on suggestions from Krippendorff and Bock (2008) and Nunnally and Bernstein (1978)

Table 4 Fit indices

	RMSEA (*p* value)	CFI	TLI	Wald test, Initial as baseline[a]	Wald test, Modified 5-Category as baseline	Wald test, 4-Category as baseline
Initial 5-category model	.07 (.00)	.98	.97			
Modified 5-category model	.05 (.31)	.99	.99	180.81 (1.00)		
4-Category model	.05 (.13)	.99	.98		1.03 (.91)	
2-Category model	.10 (.00)	.95	.95			34.60 (.00)
Full τ-equivalent model	.07 (.00)	.98	.98			89.62 (.00)

N = 726

[a] Wald tests of the hypothesis that the more parsimonious model fits as well as the more saturated model. The first number is difference between the saturated and constrained model χ^2 statistics. The number in parenthesis is the p value

improved by shifting criteria from one category to another. The next model moved three criteria: Criterion 5 (Program Design includes significant activity in service to veterans and/or military families as part of service in the primary focus area, other focus areas, or Capacity Building) was moved from Community Impact to Recruitment and Development of Volunteers, criterion 15 (Plans and infrastructure to manage project resources, both financial and in-kind, to ensure accountability and efficient and effective use of available resources) was moved from Program Management to Organization Capacity, and criterion 18 (Examples of the applicant organization's track record in managing volunteers in the primary focus area, to

include if applicable, measuring performance in the primary focus area) was moved from Organization Capacity to Program Management. This model had superior fit, but two categories, Recruitment and Development of Volunteers and Program Management, were nearly collinear, with a correlation of 0.96. Merging these categories resulted in a more parsimonious model with improved fit over the initial model and nearly identical fit to the modified 5-category model. While there were a number of high modification indices, none made theoretical sense nor would have resulted in substantial improvements in the model fit.

The correlation between two pairs of categories—Strengthening Communities with Recruitment of Volunteers/Program Management and Strengthening Communities with Organizational Capacity—was over 0.80. Three categories were merged (the 2-Category Model reported in Table 4), but the model fit was significantly worse than the 4-category model.

After identifying the 4-category model, the next step was to test whether the models were essentially τ-equivalent or congeneric. All loadings for each dimension were constrained to be equal. Doing so resulted in a model fitting slightly worse according to all three fit measures, as well as in a Wald χ^2 test. Individual categories were also tested for τ-equivalence, and in each case the model fit was inferior. This indicates that ω is the most appropriate measure for reliability.

Given the congeneric 4-category model has good fit and is superior to the other models, it is our preferred model. The factor loadings and covariances are reported in Fig. 3. All parameters are statistically significant at the 0.0001 level. These results indicate that each category is unidimensional and is generally well reflected by the criteria. The four categories are highly correlated, as shown in the path diagram in Fig. 4. Further merging categories did not result in better fit, however. Whether this detracts from the model validity is assessed in the section on discrimination below.

After establishing the dimensionality and confirming the factor structure of the model, invariance was tested across subsamples. As described above, applications were split into two groups based on two variables: the total funding in each funding opportunity (high versus low) and the focus area (Healthy Futures versus other areas). Table 5 reports the model fit statistics for each degree of measurement invariance for each grouping variable.

The results of the measurement invariance tests show that under both groupings, the model fit does not deteriorate as the level of invariance increases. Therefore, the preferred model has a high (strict) degree of measurement invariance.

The last step in establishing factorial validity is to assess the convergent and discriminate validity of the model. Table 6 provides the relevant output to assess these aspects of validity, including ω, average variance extracted, average shared variance, and maximum shared variance. For all four categories, the requirements for convergent and discriminant validity were met. ω was higher than average shared variance of the criteria in all cases. Average shared variance is higher than 0.5 in all cases. For all categories, the average variance extracted was higher than the average shared variance and the maximum shared variance of the categories, indicating that the dimensions are distinct enough from one another, despite their high correlations.

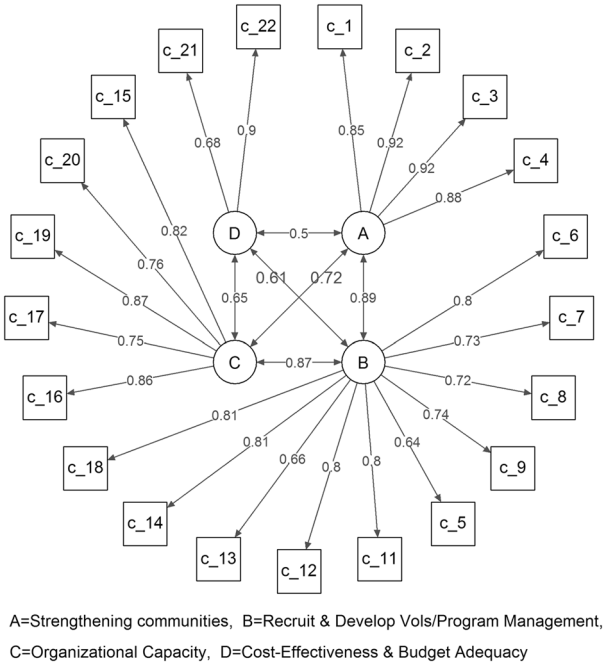

Fig. 4 Path diagram for final 4-category model. N = 726. Coefficients on paths between criteria and categories are factor loadings. Coefficients on paths among categories are correlations. All factor variances are constrained to one

Table 5 Measurement invariance fit statistics

Invariance	CFI	TLI	RMSEA
Grouping variable = high and low funding			
Configural	.99	.98	.05 (.64)
Weak	.99	.99	.05 (.81)
Strong	.99	.99	.05 (.85)
Strict	.99	.99	.05 (.94)
Grouping variable = healthy futures focus area and all other focus areas			
Configural	.99	.98	.05 (.63)
Weak	.99	.98	.05 (.80)
Strong	.98	.99	.05 (.86)
Strict	.99	.99	.04 (.97)

Measurement Reliability

The first column in Table 6 reports the reliability for each category and for the total model, as measured by ω. All categories exceeded the 0.8 threshold, with the

Table 6 Convergent and discriminate validity statistics

Category	Reliability ω	Average variance extracted	Average shared variance	Maximum shared variance
Strengthening communities	0.94	0.80	0.23	0.39
Recruitment and dev of volunteers/ program management	0.93	0.57	0.23	0.39
Organizational capacity	0.91	0.66	0.23	0.39
Effectiveness and budget adequacy	0.78	0.64	0.23	0.39
Total model	0.97			

exception of Cost Effectiveness and Budget Adequacy, which was close at 0.78. This indicates that they have sufficient measurement reliability for their intended purpose. Reliability for the entire instrument across all 4 categories was estimated at 0.97, suitable for our high-stakes purposes (Brown 1910; Fornell and Larcker 1981; Spearman 1910).

Discussion

A four-step process was presented to establish the construct validity and the reliability of grant-making decisions: determine standards for qualified applicants, assess inter-reviewer reliability, assess factorial validity, and assess measurement reliability. Given the demands for increased accountability on grant-making entities, this process can be used by others to identify the validity and reliability of grant decisions. The specific findings to this study illustrate how the process is implemented, how to interpret the results, and its limitations.

This analysis confirms that the 2013 RSVP grant competition represented a valid and reliable high-stakes test to determine funding decisions. The passing scores to determine minimally qualified and best-qualified applicants were found to be near the highest points of reliability for the rating procedure. The inter-reviewer reliability for each criterion was on average above the 0.70 threshold. Staff revised the procedures and instructions for the criteria that had low inter-reviewer reliability in the FY 2014 RSVP Notice of Funding Opportunity. Inter-reviewer reliability for the overall instrument scores was very high, above the 0.90 criterion for high-stakes purposes.

The final model contained four categories of criteria: Strengthening Communities, Recruitment and Development of Volunteers/Program Management, Organizational Capacity, and Cost Effectiveness and Budget Adequacy. The analysis merged criteria related to volunteer and program management, suggesting these are essentially the same issues for in the RSVP review process. Analysis also suggested

that three criteria were better aligned in other domains. Criterion 5 (Program design includes significant activity in service to veterans and/or military families as part of service in the primary focus area, other focus areas or capacity building) was moved from Strengthening Communities to Recruitment and Development of Volunteers in the final model. This may be because the scoring rubric instructed reviewers to give an 'Excellent' score for this criterion for applications that accounted for the 'unique value of service by RSVP volunteers who are veterans and/or military family members' (Senior Corps 2013b). Criterion 15 (Plans and infrastructure to manage project resources to ensure accountability and efficient and effective use of resources) was moved from Program Management to Organization Capacity. It is possible that issues pertaining to available resources are better aligned with capacity rather than management. Criterion 18 (Examples of the applicant organization's track record in managing volunteers in the primary focus area) was moved from Organization Capacity to Program Management in the final model. These findings were not available in time to inform the 2014 Notice of Funding Opportunity but did influence the fiscal year 2015 RSVP Notice.

Importantly, the last domain, Cost Effectiveness and Budget Adequacy, was left with only 2 criteria. The last criterion was removed from the model due to poor fit. In general, two criteria are insufficient to accurately measure the underlying construct, and this is particularly true in this case as the remaining criteria focused specifically on volunteer expenses. In formal feedback during the review, the application reviewers expressed concerns that these selection criteria did not fully account for all aspects of the domain. In addition, these same criteria were found to have lower inter-reviewer reliability than the other criteria. In order to address these concerns, it was decided that for the 2014 Notice of Funding Opportunity, the criteria for Cost Effectiveness and Budget would be reviewed exclusively by financial management staff with expertise in this specific area, rather than program staff and external reviewers.

The final model was found to have strict measurement invariance, meaning that it is not biased across different subgroups. It was also found to have high convergent validity, meaning that each domain is well explained by its respective criteria, and high discriminant validity, meaning that the domains were sufficiently distinct from one another. The final grant scores had high measurement reliability, both at the category level and the overall instrument level, giving there is high confidence that the model approximated the 'true' quality of applicants.

This analysis had several important limitations. All measures of inter-reviewer reliability assume that reviewers assign ratings independent of one another. In the case of the RSVP review process, reviewers in each panel discussed the applications under review and came to a common understanding of each application's content. They were not instructed to come to a consensus when scoring applications but it is possible that some consensus did arise, which would cause the reliability measures to be biased upward (meaning they are higher than they should be). An additional limitation is that the analysis is dependent on the reviewers and applications for the 2013 competition, as well as the conditions of the rating. The publication of this analysis in no way indicates that future applications for Senior Corps or other CNCS funding opportunities will be

reviewed in a similar manner. Publication of this manuscript in no way represents an agency commitment to conduct or publish similar analysis of CNCS competitive selection processes or a change on its policy about releasing pre-decisional grant competition material. Finally, the 2013 Notice was designed to reflect both the requirements of the Serve America Act and the 42-year history of the RSVP program. Many of the particular findings of this analysis may not apply to other federal funding opportunities.

Although the objectives of this study were not to identify the reasons for the validity and reliability of the grant-making process, we can offer several hypotheses. We believe the strength our findings are due in large part to the measurement instruments and procedures on the one hand, and applicant understanding of these procedures on the other. The instruments refer the rating forms containing the selection criteria and scoring rubrics used by application reviewers. The rating forms were heavily vetted by subject matter experts throughout the agency, incorporating feedback from staff that monitor grantees and review applications, leadership who develop policies, and the research and evaluation office that assesses performance. The project team developing the materials worked diligently to ensure that the selection criteria both aligned with the goals of RSVP, and represented the concepts that could provide decision-makers the right information needed to determine who to fund. The review procedures were highly standardized, with training and assistance provided to reviewers to clarify questions on the criteria, and quality control procedures to ensure ratings were sufficiently supported by narrative statements. All of these processes helped increase the likelihood that any differences in reviewer ratings were due to their own knowledge and true differences in applications, rather than reviewer understanding or interpretation of criteria.

Finally, the rating form, selection criteria, and scoring rubrics were made available to all applicants along with the funding announcement. This helped applicants focus on what they would be measured against, increasing the likelihood that differences in ratings were due to actual differences in applicant quality rather than applicant interpretation of the criteria.

Establishing the validity and reliability of the grant-making process should be an important component of a fair and transparent grant-making process. The processes outlined in this study—well established in other fields—can be applied to other grant-making contexts, similarly providing insight into improved, more defensible grant-making decisions. In addition, from 2012 to 2014, the Obama Administration released proposed budgets for CNCS for fiscal years 2013, 2014, and 2015 that include proposals to expand competition to two other Senior Corps programs: the Senior Companion Program and the Foster Grandparent Program. This analysis of the 2013 RSVP competition should assure the public of the capacity of CNCS to compete all Senior Corps grants.

Appendix

See Tables 7 and 8.

Table 7 Selection criteria definitions

Category and sub-category	Criterion	Description
Program design		
Strengthening communities	1	Demonstrates that community needs identified in the Primary Focus Area exist in the geographic service area and are currently unmet
	2	Demonstrates plans and infrastructure to manage RSVP volunteers and their stations as a highly effective means to addressing the identified community need(s) in the Primary Focus Area
	3	Describes how the service activities in the Primary Focus Area lead to National Performance Measure outputs or outcomes
	4	Connects the following three major elements in the Primary Focus Area to each other: i. The community needs identified; ii. The service activities that will be carried out by RSVP volunteers; and iii. The anticipated National Performance Measure output(s) or, if possible, National Performance Measure outcome(s)
	5	Has a Program Design that includes significant activity in service to veterans and/or military families as part of service in the Primary Focus Area, Other Focus Areas, or Capacity Building
Recruitment and development of volunteers	6	Plan and infrastructure to create high-quality RSVP volunteer assignments with opportunities such as share their experiences, abilities, and skills to improve their communities and themselves through service in their communities
	7	Plan and infrastructure to ensure RSVP volunteers receive training needed to be effective in their assignments
	8	Plan and infrastructure to recruit an RSVP volunteer pool from one of following populations: Individuals of all races, ethnicities, sexual orientation, and degrees of English language proficiency Veterans and military family members as RSVP volunteers RSVP volunteers with disabilities, including individuals with age-related disabilities
	9	Plan and infrastructure to retain and recognize and appreciate RSVP volunteers
National performance measure work plan	10	In assessing the work plans, applications will receive credit for a percentage of unduplicated RSVP volunteers in National Performance Measure outcome work plans above the minimum 10 %
Organizational capacity		

Table 7 continued

Category and sub-category	Criterion	Description
Program management	11	Plan and infrastructure to ensure management of volunteer stations in compliance with RSVP program regulations including preventing or identifying prohibited activities
	12	Plan and infrastructure to develop and/or oversee volunteer stations that address specified community needs outside the Primary Focus Area
	13	Plan and infrastructure to responsibly graduate volunteer stations to meet changing community needs, and do so in a way that minimizes disruptions to current volunteers where possible
	14	Plan and infrastructure to assure that national performance measure outcomes and outputs are measured and collected
	15	Plan and infrastructure to manage project resources, both financial and in-kind, to ensure accountability and efficient and effective use of available resources
Organizational capability	16	Plans and infrastructure to provide sound programmatic and fiscal oversight, day-to-day operational support and data collection, and clearly defined internal policies
	17	Descriptions of clearly defined paid staff positions, including how these positions will be sustained and (as applicable) identification of current staff assigned to the project
	18	Examples of the sponsor organization's track record in managing volunteers in the Primary Focus Area, to include if applicable, measuring performance in the Primary Focus Area
	19	Strong organizational infrastructure, including
		i. Tangible assets such as facilities, equipment, supplies
		ii. Governance structure and operations such as internal policies, purchasing procedures, and personnel management
		iii. Role of a community participation group, such as an RSVP Advisory Council[a], to ensure input from the community
		iv. Robust financial management systems and past experience managing federal grant funds
	20	Demonstrates the adequacy and sustainability of the applicant's proposed required non-federal financial share

Table 7 continued

Category and sub-category	Criterion	Description
Cost effectiveness/budget adequacy	21	Plan and infrastructure to provide applicable costs and reimbursable expenses to volunteers such as transportation, meals, and insurance, as well as plans and infrastructure to provide criminal history background checks as appropriate
	22	The adequacy and reasonableness of the budget to support RSVP volunteer recruitment, support, and recognition
	23	The adequacy and reasonableness of required non-federal funds budgeted

[a] Advisory Council: RSVP Federal Regulation §2553.24 requires grantees to secure community participation in local project operation by establishing an Advisory Council or a similar organizational structure with a membership that includes people knowledgeable about human and social needs of the community; competent in the field of community service and volunteerism; capable of helping the sponsor meet its administrative and program responsibilities including fund-raising, publicity and programing for impact; with an interest in and knowledge of the capability of older adults; and of a diverse composition that reflects the demographics of the service area

Table 8 Descriptive statistics for selection criteria

Criteria	Mean	St. Dev.	Mode
q_1	1.7	0.9	2
q_2	1.6	0.8	2
q_3	1.5	0.8	2
q_4	1.6	0.9	2
q_5	1.2	0.9	1
q_6	1.6	0.8	2
q_7	1.5	0.9	2
q_8	1.3	0.9	1
q_9	1.6	0.8	2
q_10	1	1.3	1
q_11	1.5	0.9	2
q_12	1.4	0.9	2
q_13	1.3	1	2
q_14	1.5	0.8	2
q_15	1.8	0.8	2
q_16	1.7	0.7	2
q_17	1.6	0.7	2
q_18	1.6	0.7	2
q_19	1.8	0.8	2
q_20	1.6	0.8	2
q_21	1.4	0.8	1
q_22	1.4	0.7	1
q_23	0.7	0.9	0

All criteria were rated on a 0–3 scale, where 0 = *does not meet*, 1 = *fair*, 2 = *good*, and 3 = *excellent*

References

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling: A Multidisciplinary Journal, 13*(2), 186–203.

Bond, T. G., & Fox, C. M. (2013). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). New York: Psychology Press.

Bornmann, L., Mutz, R., & Daniel, H.-D. (2010). A reliability-generalization study of journal peer reviews: A multilevel meta-analysis of inter-rater reliability and its determinants. *PLoS ONE, 5*(12), e14331.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 1904–1920, 3*(3), 296–322.

Chandra, A. (2011). *Building community resilience to disasters: A way forward to enhance national health security*. Santa Monica, CA: Rand Corporation.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests* (Vol. xv). Thousand Oaks, CA: Sage Publications Ltd.

CNCS. (2012). Funding Available to Engage Older Americans in High-Impact Service. Corporation for National and Community Service. Retrieved August 2, 2012 from http://www.nationalservice.gov/newsroom/press-releases/2012/funding-available-engage-older-americans-high-impact-service

CNCS. (2013a). Congressional Budget Justification Fiscal Year 2014. Corporation for National and Community Service.

CNCS. (2013b). The Corporation for National and Community Service Competitive Review and Selection Process, 2013. Corporation for National and Community Service. Retrieved from http://www.nationalservice.gov/sites/default/files/documents/cncs_2013_grant_review_and_selection_process_0.pdf

CNCS. (2013c). National Service Agency Grants to Support 80,000 Senior Volunteers. Press Release, Corporation for National and Community Service. Retrieved February 22, 2013 from http://www.nationalservice.gov/newsroom/press-releases/2013/national-service-agency-grants-support-80000-senior-volunteers.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334.

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research, 18*(1), 39–50.

Gignac, G. E. (2014). On the inappropriateness of using items to calculate total scale score reliability via coefficient alpha for multidimensional scales. *European Journal of Psychological Assessment, 30*(2), 130–139.

Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability what they are and how to use them. *Educational and Psychological Measurement, 66*(6), 930–944.

Hair, J. F., Anderson, R. E., Tatham, R. L., & William, C. (2010). *Multivariate data analysis* (7th ed.). Upper Saddle River, NJ: Prentice-Hall Inc.

Haynes, S. N., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment, 7*(3), 238–247.

Kotchen, T. A., Lindquist, T., Malik, K., & Ehrenfeld, E. (2004). NIH peer review of grant applications for clinical research. *JAMA, 291*(7), 836–843.

Krippendorff, K., & Bock, M. A. (2008). Testing the reliability of content analysis data: What is involved and why. *The Content Analysis Reader*. Thousand Oaks, CA: SAGE.

Lord, F. M., Novick, M. R., & Birnbaum, A. (1968). *Statistical theories of mental test scores*. Oxford: Addison-Wesley.

Marsh, H. W., Jayasinghe, U. W., & Bond, N. W. (2008). Improving the peer-review process for grant applications: Reliability, validity, bias, and generalizability. *American Psychologist, 63*(3), 160–168.

McDonald, R. P. (1999). *Test theory: A unified treatment*. New York: Psychology Press.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*(4), 525–543.

Meyer, P. (2010). *Understanding measurement: Reliability*. New York: Oxford University Press.

Muthén, B., & Asparouhov, T. (2005). Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus. *Mplus Web Notes, 4*(5), 1–22.

Mutz, R., Bornmann, L., & Daniel, H.-D. (2012). Heterogeneity of inter-rater reliabilities of grant peer reviews and its determinants: A general estimating equations approach. *PLoS ONE, 7*(10), e48509.

Nunnally, J. C., & Bernstein, I. H. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Raykov, T. (2001). Bias of coefficient a for fixed congeneric measures with correlated errors. *Applied Psychological Measurement, 25*(1), 69–76. doi:10.1177/01466216010251005.

Revelle, W., & Zinbarg, R. E. (2008). Coefficients alpha, beta, omega, and the GLB: Comments on sijtsma. *Psychometrika*, *74*(1), 145–154.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 186–203.

Rothwell, P. M., & Martyn, C. N. (2000). Reproducibility of peer review in clinical neuroscience Is agreement between reviewers any greater than would be expected by chance alone? *Brain, 123*(9), 1964–1969. doi:10.1093/brain/123.9.1964.

Senior Corps. (2010). Staff Webinar: Presenting the RSVP Re-Competition Webinar to RSVP Grantees. Retrieved May 2010 from https://www.nationalserviceresources.gov/files/state-staff-webinar.ppt

Senior Corps. (2012). Announcement of Federal Funding Opportunity: 2013 RSVP Competition. Corporation for National and Community Service.

Senior Corps. (2013a). 2013 RSVP External Blended Review Participants. Corporation for National and Community Service. Retrieved from http://www.nationalservice.gov/sites/default/files/upload/FY2013 percent20RSVP percent20External percent20Reviewers.508.pdf.

Senior Corps. (2013b). Announcement of Federal Funding Opportunity: 2014 RSVP Competition. Corporation for National and Community Service. Retrieved from http://www.nationalservice.gov/build-your-capacity/grants/funding-opportunities/2013/2014-rsvp-competition

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology 1904-1920, 3*(3), 271–295.

Stemler, S. (2014). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. Stemler, Steven E. Practical Assessment, Research and Evaluation, *9*(4), 1–19. Retrieved from http://www.pareonline.net/getvn.asp?v=9andn=4

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52*(3), 393–408.

Vandenberg, R. J. (2002). Toward a further understanding of and improvement in measurement invariance methods and procedures. *Organizational Research Methods, 5*(2), 139–158.