Published in final edited form as:

J Health Psychol. 2016 September ; 21(9): 2033–2041. doi:10.1177/1359105315569859.

The reliability of psychological instruments in community samples: A cautionary note

James A Shepperd, Amber S Emanuel, Virginia J Dodd, and Henrietta L Logan

University of Florida, USA

Abstract

Although researchers norm and validate measures of psychological constructs largely on educated samples, they often use these instruments more broadly, assuming generalizability. We examined whether the assumption of generalizability is warranted. We administered three commonly used psychological measures—the Behavioral Activation/Behavioral Inhibition Scale, the Regulatory Focus Questionnaire, and the Need for Cognition Scale—to a community sample ($N = 332$) with limited education. For the three instruments, five of seven scales/subscales had unsatisfactory reliabilities. Internal consistency was lowest among participants with less education. The results suggest that instruments normed on educated samples may not generalize to uneducated samples.

Keywords

cognitive ability; community psychology; individual differences; reliability; sampling

Measures of psychological instruments are typically designed by researchers who carefully develop, test, and subject items to rigorous psychometric analyses to assure that the end measures are internally consistent, reliable across time, and demonstrate strong convergent, discriminant, and predictive validity. Most researchers readily acknowledge, and decades of research comparing culturally different populations verify that psychological instruments often generalize poorly across cultures (Heine et al., 2002; Taras et al., 2009). Differences in language and cultural practices between populations reveal the need to adapt instruments to specific populations (Chen et al., 1995; Gjersing et al., 2010; Tucker et al., 2006). Surprisingly, little research has examined the presence and meaning of population-specific, within-group differences in characteristics such as education or economic status. The diversity within communities and the need to tailor interventions to specific strata within a population make attention to these within-group differences critical.

For example, instruments validated and normed on college students or other academically skilled samples are often administered to quite different groups, perhaps with the assumption that people are largely interchangeable and that an instrument normed in one group is generally appropriate for anyone conversant in the language of the instrument. Investigators may make exceptions for people who are mentally challenged, but instrument developers

Reprints and permissions: sagepub.co.uk/journalsPermissions.nav

Corresponding author: James A Shepperd, Department of Psychology, University of Florida, P.O. Box 112250, Gainesville, FL 32611-2250, USA. shepperd@ufl.edu.

rarely specify boundaries for appropriate use. The central question is whether investigators can assume generality. Are measures of psychological instruments normed and validated in one group, appropriate for other groups within the same, larger culture that may differ in language skills or in their interpretation of the instrument items?

This article provides an initial test of a hypothesis that appears generally ignored or overlooked by investigators conducting research in the community: psychological instruments that are developed on and perform well in educated, highly literate samples, may be unreliable in less educated samples. This hypothesis arises from several observations. First, the reading skills required by many psychological instruments exceed the ability of people with low literacy skills, leaving such participants unable to understand the meaning of the items and requiring more cognitive effort than they may be able or willing to put forth.

Second, different people may interpret the same items quite differently. For example, the meaning of many words is culturally based and the meaning intended during scale development and shared by an educated norming group may not be shared by people from other groups (Dodd et al., 2012). Relatedly, the items that describe events or ideas, or that use colloquialism relevant to people in mainstream America with at least some college education, may have little meaning to people outside the mainstream. For example, the Regulatory Focus Questionnaire (RFQ)—an instrument examined in the present research—has an item that reads, “How often have you accomplished things that got you ‘psyched’ to work even harder?” (Higgins et al., 2001). This colloquialism may confuse participants, or the word “psyched” may evoke a negative image.

Third, instruments developed on groups such as college students may reference activities or behaviors that have little relevance or meaning to non-college students or to people within the same, larger culture who have different experiences. For example, many items from the RFQ make references to participants’ behavior around their parents, which may have little relevance to people not raised by their parents or who have a different relationship with their parents than do people from the norming sample.

Fourth, many researchers assume the constructs they are measuring are universally relevant and meaningful to all people. People differ only in where they fall along the continuum representing the construct. An alternative perspective, grounded in the ideographic approach to measurement, assumes that all traits are not equally relevant to all people (Allport, 1937). Specifically, people differ in whether a given trait affects their cognitions and behavior. If a trait is less relevant to a person, then the person is contributing more error to the measurement process. If a trait is more relevant to a person, then the person is contributing less error to the measurement process. It is possible that the constructs assessed by instruments normed on educated, highly literate samples are less relevant to less educated, less literate samples. In terms of classical test theory (Spearman, 1907), these instruments would show low reliability because the responses of less educated samples largely represent error variance rather than true score (Britt and Shepperd, 1999).

This study was part of a project that examined whether three commonly used psychological instruments—the Behavioral Inhibition Scale/Behavioral Activation Scale (Carver and White, 1994), RFQ (Higgins et al., 2001), and Need for Cognition Scale (Cacioppo et al., 1984)—predict responses to health messages about oral cancer. A key component of the research was to verify that the reliability and validity of these instruments in our sample of rural Black Americans. When we discovered that these measures showed poor reliability, we formulated and tested the hypothesis that one cause was low literacy/education, and that these scales would show low reliability coefficients particularly among the less educated participants.

Methods

Sample

Participants were 332 people taking part in IRB approved research on oral cancer screening among rural residents of north central Florida (see Table 1 for sample characteristics). We recruited participants by posting flyers in local businesses, through community events held in the county, and through a snowball technique where participants were encouraged to alert friends about the research. After a researcher consented them, participants provided demographic information and completed three commonly used psychological instruments.

Instruments

Behavioral Inhibition Scale/Behavioral Activation Scale (BIS/BAS)—The BIS/BAS (Carver and White, 1994) is a 20-item instrument with four subscales. One subscale measures behavioral inhibition (7 items), which assesses anticipation of punishment. The remaining subscales measure behavioral activation. The BAS-fun scale (4 items) measures desire for new rewards, the BAS-Reward scale (5 items) measures anticipation of reward, and the BAS-Drive scale (4 items) measures pursuit of desired goals. The BIS/BAS uses a 4-point scale anchored by 1 = *very true for me* and 4 = *very false for me*. Example items include, “Criticism or scolding hurts me quite a bit” (BIS), “I’m always willing to try something new if I think it will be fun” (BAS-Fun), “When I’m doing well at something, I love to keep at it” (BAS-Reward), and “I go out of my way to get things I want” (BAS-Drive).

The authors of the BIS/BAS report reliabilities coefficients of $\alpha = .74$ for the BIS and reliability coefficients of $\alpha = .66$ to $.73$ for the BAS (Carver and White, 1994). A review of the literature for this article that paired the search terms *BIS/BAS scale* and *community sample* uncovered 15 articles. Ten reported no reliability information. Of the five articles that did report reliability information separately by scale, Cronbach’s alpha—a measure of internal consistency—ranged from $.72$ to $.80$ for the BIS and from $.62$ to $.84$ for the various subscales of the BAS (Cherbuin et al., 2008; Hall et al., 2008).

Regulatory Focus Questionnaire (RFQ)—The RFQ (Higgins et al., 2001) is an 11-item instrument with two subscales. The promotion scale (6 items) measures concern with achieving success and the presence of positive outcomes or gains (e.g. “Compared to most people, are you typically unable to get what you want out of life?”). The prevention scale (5

items) measures concerns with avoiding failures and the presence of negative outcomes or losses (e.g. “Growing up, did you ever act in ways that your parents thought were objectionable?”). The RFQ uses 5-point scales with varying anchors (e.g. 1 = *never or seldom/never true/certainly false* and 5 = *very often/very often true/certainly true*).

The authors of the RFQ report reliability coefficients of $\alpha = .80$ for the Prevention Scale and $\alpha = .73$ for the Promotion Scale (Higgins et al., 2001). A review of the literature for this article that paired the search terms *self-regulatory focus* and *community sample* uncovered six articles. Two did not report reliability information. Of the four that did, Cronbach’s alphas ranged from .53 to .71 for promotion scale and from .70 to .86 for the prevention scale. Consistent with the guiding hypothesis, Cronbach’s alpha was lowest for a sample with low literacy skills (Martinez et al., 2013) and highest for a highly educated sample (Worth et al., 2005) and another sample that was likely highly functioning because participants were recruited via mTurk and thus had access to computers and used the Internet (Joel et al., 2014).

Need for Cognition Scale—The Need for Cognition Scale (NCS; Cacioppo et al., 1984) is an 18-item instrument that measures whether people are motivated to engage in challenging cognitive activities. The NCS uses a 9-point scale anchored by $-4 = \text{very strongly disagree}$ and $+4 = \text{very strongly agree}$. An example item is, “I find satisfaction in deliberating hard for long hours.” Due to an error we made in labeling the endpoints, we had useable data from only 65 participants for the NCS.

The authors of the NCS report a reliability coefficient of $\alpha = .90$ for the scale (Cacioppo et al., 1996). A review of the literature for this article that paired the search terms *need for cognition* and *community sample* uncovered 20 articles. Of the 10 articles that reported reliability information, Cronbach’s alpha ranged from .57 for a sample of substance users on probation (Czuchry and Dansereau, 2004) to .89 for a sample of Amish adults (McGuigan and Scholl, 2007).

Procedures

Participants completed all measures individually via paper and pencil. A community coordinator arranged data collection efforts and assigned data collectors to community events as needed. The data collectors instructed participants to complete the questions to the best of their ability. Participants completed the instruments in no particular order.

Data analyses

We computed reliability using Cronbach’s alpha, and compared high and low education groups using the Feldt test except for the NCS, where the small sample size necessitated using the Fisher–Bonett test (Kim and Feldt, 2008). Cronbach’s alpha is a function of the average inter-item correlation (AIC) within a scale and the number of scale items. However, the number of items in several of the scales was low. Thus, we computed the AIC for the entire sample and for the high and low education groups. We calculated Flesch–Kincaid Grade level and the Flesch reading ease values for each instrument using a feature of Microsoft Word. We conducted all analyses with Stata or with SAS.

Results

Table 2 presents Cronbach's alpha for each instrument for the entire sample and separately for participants by education level: high school degree or less ($n = 225$) versus more than a high school degree ($n = 88$). We chose to dichotomize the groups into two groups based on the recognition that *social promotion* was common in the past in some school system and grade completed was not always consistent with literacy skills (Riley, 1999). For the NCS, 40 participants had a high school degree or less, 20 had more than a high school education, and 5 did not report their education. As evident in the first column of numbers, Cronbach's alpha was low in all instances. More striking is the difference in reliabilities between participants categorized as low versus high in education. Cronbach's alpha was consistently greater in the high than in the low education group, and significantly so in four of seven cases.¹ Importantly, Cronbach's alpha remained low even in the high education group.

Table 2 presents the AIC for the entire sample and for the high and low education groups. Looking first at the entire sample, the AICs were generally low, falling below .20 in four of seven instances. Although the scale authors did not report the AICs for their instruments, we obtained unpublished AICs for the RFQ and BAS (but not the NCS) from samples of undergraduate students (Emanuel, 2014). The AICs were consistently higher in the undergraduate samples. Specifically, in a sample of 126 undergraduates, the AIC was .17 for the promotion scale and .64 for the prevention scale. In another sample of 81 undergraduates, the AIC was .37 for the BIS scale, .44 for the BAS-fun scale, .27 for the BAS reward scale, and .44 for the BAS drive scale.

When we separated participants into high and low education groups, the AIC exceeded .20 in four of seven instances among high education participants, but only in two of seven instances among low education participants. In all cases, low education participants had lower average AICs than high education participants, suggesting that the items within scales are weakly correlated at best. The difference between education groups in the average AIC was significant for the NCS, two RFQ scales and the BIS subscale, $p's < .01$, $d's > .60$.

Finally, the last two columns of Table 2 present the Flesch–Kincaid Grade level and the Flesch reading ease for each instrument. The Flesch–Kincaid Grade level is an estimate of the grade level a participant would need to have completed to successfully comprehend the text. The Flesch Reading Ease is an estimate of how easily text can be comprehended based on sentence and word length (Flesch, 1948). Higher scores indicate that the material is easier to read. Both measures are indicators of reading difficulty. Although results from both indicators suggest that all the instruments are comprehensible for someone with a high school degree, many of the participants had less than a high school degree. As evident from Table 2, the instrument with the lowest reliability and the lowest mean inter-item correlation—the Regulatory Focus Promotion Scale—was among the most difficult to read.

One last point deserves mention. As is typical with studies involving questionnaire responses, participants occasionally skip items. We found that 102 of our 332 participants

¹Further subdividing the low education group into participants with versus without a high school degree did not yield meaningful differences because of low statistical power.

(30%) skipped or otherwise failed to answer one or more items on the psychological instruments we examined. The correlation between education level and completing all items was $r = -.15$, $p = .006$. Although the correlation is small, an obvious reason why less educated participants skipped items is that they did not understand the meaning and thus were reluctant to provide a response. This finding is consistent with our central point that these commonly used instruments may be problematic for less educated groups.

Discussion

Consistent with the hypothesis, several instruments that appear reliable in educated, highly literate norming samples were less reliable in the less educated, less literate sample. The low reliability appears to arise at least in part from the instruments requiring education and reading comprehension skills that exceeded those of the sample. Specifically, the scale reliabilities were particularly low among the less educated participants, and the two instruments with the lowest reliability and the lowest AIC—the NCS and the Regulatory Focus Promotion scale—were among the most difficult to read.

Several of the instruments performed poorly even among the high education group. Yet the high education group was high only in a relative sense. Conversations with participations revealed that many of the post-high school education participants interpreted “some college” as attending a trade school (e.g. beautician school) or taking a few courses at a community college. A 4-year baccalaureate degree, or even past enrollment at 4-year degree institution, was rare in the sample. Moreover, other research conducted in this population reveals that these participants had low literacy skills and were infrequent readers (Shepperd et al., 2014).

Not all instruments performed poorly. The reliability coefficients for the three BAS scales were generally consistent with the normative data presented by the instrument developers. The three BAS scales may have demonstrated higher reliability because they tapped constructs that were more universal or because the item wording was simpler.

Implications and limitations

Researchers have recently raised concerns about the generality of research findings and theories based on educated, middle-class participants who are mostly White and affluent, and from western, industrialized, democratic countries (Henrich et al., 2010). The current findings extend these concerns, suggesting that some psychological instruments may not generalize beyond the samples with which they were developed. Stated bluntly, instruments developed on college students or other affluent samples may translate poorly to less educated, less affluent populations. The opposite may also be true; instruments developed on less educated, less affluent populations may not generalize to college students or other affluent samples. These findings speak to the need for diverse samples during the development of instruments. Homogeneous samples during scale development can handicap future research and undermine future studies and the information gleaned.

Although this point may hardly seem new, researchers are using these scales in community settings where the sample in some instances differs markedly from the norming sample. The widespread, indiscriminate use of psychological instruments likely extends beyond the

scales examined here. Apparently, the limitations of generality of instruments normed on one group such as college students, although likely taught in every graduate measurement class, are not taken seriously. The current findings are also a reminder that people within a heterogeneous culture such as the United States may respond quite differently to the same instrument.

And there are other implications. Perhaps one reason some psychological interventions fail to produce changes in behavior, or show weak or confusing results, is not because the interventions are misguided. Rather, the messages researchers create or the psychological instruments researchers use may be unreliable or lack meaning in the target population. Three examples illustrate this point. First, the need emerged early in our research to use the term *mouth and throat* cancer rather than *oral* cancer in message campaigns. As various participants commented, “I don’t have an oral, I have a mouth,” and “Oral, is that the same as the anus?” Words that have clear meaning for one group of people may have no meaning or a different meaning for another group. Second, health researchers often tailor interventions for specific subgroups. For example, they may tailor some messages to people who are high in the need for cognition and other messages to people who are low in the need for cognition. Interventions using message tailoring may be ineffective if the instruments used to define groups are unreliable in the target group and thus do not actually distinguish between people who are high versus low on the underlying psychological construct. Third, researchers often include measures of attitudes, cognition, and other psychological constructs to test potential moderators and mediators of the effects they observe. These tests may fail if the measures are unreliable in the target population.

The present data have limitations. Most notably, this community sample of mostly Black participants differs from the typical norming sample, often White college students attending large universities, in many ways. It is unclear what aspect of the sample was responsible for the difference between the present results and the results of studies that used other samples. Importantly, this limitation serves to highlight a central message of this article: Samples from the same, larger culture (the United States of America) can differ in many ways, and the differences can produce variations in response to a given psychological instrument.

A second limitation is that it is unknown whether the low reliability of these measures in this sample would impair their predictive validity. However, given that high reliability is necessary for validity (Anastasi, 1982), it seems likely that the predictive validity would also be low. Third, this study only examined responses of residents of one rural community in north central Florida. The pervasiveness of the reliability problem described here remains unknown. Fourth, it remains unknown whether the problems with internal consistency observed in samples with limited education are unique to these instruments or generalize to other instruments. Finally, item response theory (Lord, 1980) is useful for examining group differences in response to specific items. However, the procedures require a larger sample than reported here.

Recommendations

Our observations lead to several recommendations. First, researchers must be sensitive to the psychometric limitations of their instruments and take seriously the admonition that an

instrument normed in one demographic group may not be appropriate for another demographic group. Researchers should test their instruments in the target population before administration. Second, researchers should be cognizant of the assumptions implicit in items from the instruments they wish to use, such as that participants are raised by their parents or that their parents are alive. Moreover, some instruments describe activities, thoughts, or beliefs that, although pertinent to groups such as college students, may be irrelevant to other groups within the same larger culture. Our findings suggest that community samples can differ from the norming sample in important ways, and researchers ignore these differences at their peril.

Acknowledgements

The authors thank Yi Guo and Greg Webster for assistance in data analysis.

Funding

This work was supported by the Southeast Center for Research to Reduce Disparities in Oral Health 1U54DE019261-01 H. Logan PI.

References

- Allport, GW. *Personality: A Psychological Interpretation*. New York: Henry Holt and Company; 1937.
- Anastasi, A. *Psychological Testing*. 5th edn. New York: Macmillan; 1982.
- Britt TW, Shepperd JA. Trait relevance and trait assessment. *Personality and Social Psychology Review*. 1999; 3(2):108–122.
- Cacioppo JT, Petty RE, Kao CF. The efficient assessment of need for cognition. *Journal of Personality Assessment*. 1984; 48(3):306–307. [PubMed: 16367530]
- Cacioppo JT, Petty RE, Feinstein JA, et al. Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. *Psychological Bulletin*. 1996; 119(2):197–253.
- Carver CS, White TL. Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: The BIS/BAS scales. *Journal of Personality and Social Psychology*. 1994; 67:319–333.
- Chen C, Lee S, Stevenson H. Response style and cross-cultural comparisons of rating scales among East Asian and North American students. *Psychological Science*. 1995; 6(3):170–175.
- Cherbuin N, Windsor TD, Anstey KJ, et al. Hippocampal volume is positively associated with behavioural inhibition (BIS) in a large community-based sample of mid-life adults: The PATH through life study. *Social Cognitive and Affective Neuroscience*. 2008; 3(3):262–269. [PubMed: 19015118]
- Czuchry M, Dansereau DF. The importance of need for cognition and educational experience in enhanced and standard substance abuse treatment. *Journal of Psychoactive Drugs*. 2004; 36(2):243–251. [PubMed: 15369206]
- Dodd VJ, Riley JL 3rd, Logan HL. Developing an oropharyngeal cancer (OPC) knowledge and behaviors survey. *American Journal of Health Behavior*. 2012; 36(5):589–601. [PubMed: 22584087]
- Emanuel, A. Average inter-item correlations for the regulatory fit and the behavioral activation and behavioral inhibition scales. Kent, OH: Kent State University; 2014. Unpublished data
- Flesch R. A new readability yardstick. *Journal of Applied Psychology*. 1948; 32(3):221–233. [PubMed: 18867058]
- Gjersing L, Caplehorn JR, Clausen T. Cross-cultural adaptation of research instruments: Language, setting, time and statistical considerations. *BMC Medical Research Methodology*. 2010; 10 Article 13.

- Hall P, Coons M, Vallis T. Anxious temperament and disease progression at diagnosis: The case of Type 2 diabetes. *Psychosomatic Medicine*. 2008; 70(7):837–843. [PubMed: 18606721]
- Heine S, Lehman D, Peng K. What's wrong with cross-cultural comparisons of subjective Likert Scales?: The reference-group effect. *Journal of Personality and Social Psychology*. 2002; 82(6): 903–918. [PubMed: 12051579]
- Henrich J, Heine SJ, Norenzayan A. The weirdest people in the world? *Behavioral and Brain Sciences*. 2010; 33(2–3):61–83. [PubMed: 20550733]
- Higgins ET, Friedman RS, Harlow RE, et al. Achievement orientations from subjective histories of success: Promotion pride versus prevention pride. *European Journal of Social Psychology*. 2001; 31(1):3–23.
- Joel S, Burton CM, Plaks JE. Conservatives anticipate and experience stronger emotional reactions to negative outcomes. *Journal of Personality*. 2014; 82(1):32–43. [PubMed: 23336710]
- Kim S, Feldt LS. A comparison of tests for equality of two or more independent alpha coefficients. *Journal of Educational Measurement*. 2008; 45(2):179–193.
- Lord, FM. *Applications of Item Response Theory to Practical Testing Problems*. Mahwah, NJ: Lawrence Erlbaum Associates; 1980.
- McGuigan WM, Scholl C. The effect of contact on attitudes toward old order Amish. *Journal of Applied Social Psychology*. 2007; 37(11):2642–2659.
- Martinez JL, Duncan LR, Rivers SE, et al. Examining the use of message tailoring to promote physical activity among medically underserved adults. *Journal of Health Psychology*. 2013; 18(4):470–476. [PubMed: 22689590]
- Riley, R. *Taking Responsibility for Ending Social Promotion: A Guide for Educators and State and Local Leaders*. Washington, DC: U.S. Department of Education; 1999.
- Shepperd JA, Howell JL, Logan H. A survey of barriers to screening for oral cancer among rural Black Americans. *Psycho-Oncology*. 2014; 23(3):276–282. [PubMed: 24115516]
- Spearman C. Demonstration of formulæ for true measurement of correlation. *American Journal of Psychology*. 1907; 18:161–169.
- Taras V, Rowney J, Steel P. Half a century of measuring culture: Review of approaches, challenges, and limitations based on the analysis of 121 instruments for quantifying culture. *Journal of International Management*. 2009; 15(4):357–373.
- Tucker K, Ozer D, Lyubomirsky S, et al. Testing for measurement invariance in the Satisfaction with Life Scale: A comparison of Russians and North Americans. *Social Indicators Research*. 2006; 78:341–360.
- Worth KA, Sullivan HW, Hertel AW, et al. Avoidance goals can be beneficial: A look at smoking cessation. *Basic and Applied Social Psychology*. 2005; 27(2):107–116.

Table 1

Sample characteristics.

Characteristic	<i>n</i>
Sex	
Male	234
Female	89
Did not respond	9
Marital status	
Single/cohabitating	191
Married	103
Divorced/widowed	32
Did not respond	6
Education	
Less than HS	60
HS (or GED)	165
College or more	88
Did not respond	19
Race	
Black	315
Native American	1
White	4
Other	5
Did not respond	7
Age range	
18–24	48
25–39	103
40–50	85
51–60	50
>60	41
Did not respond	5

Table 2

Reliabilities, mean item correlations, and reading level for the scales.

Instrument (number of items)	Cronbach's alpha			Average inter-item correlation			Flesch- Kincaid grade level	Flesch reading ease
	Entire sample	Low education	High education	W	Entire sample	Low education	High education	
NFC (18)	.40	.32	.70 ^a	1.96 ^a	.06	.03	.07	68.9
RFQ-Promotion (6)	.16	-.13	.51 ^a	.50	.03	-.02	.14	73.3
RFQ-Prevention (5)	.47	.41	.64 ^a	.60	.17	.14	.27	73.2
BIS (7)	.40	.27	.61 ^a	.53	.09	.05	.18	69.0
BAS-Fun (4)	.51	.48	.55	.87	.21	.19	.23	73.0
BAS-Reward (5)	.64	.63	.64	.97	.26	.25	.27	67.7
BAS-Drive (4)	.62	.62	.64	.95	.30	.30	.32	80.3

The symbol “^a” indicates that Cronbach's alpha for the two education levels differ at $p < .05$. The NFC (Need for Cognition) scale uses a 9-point response option anchored by $-4 = \text{very strongly disagree}$ and $+4 = \text{very strongly agree}$. The RFQ (Regulatory Focus Questionnaire) scales use a 5-point response option with varying anchors (e.g. $1 = \text{never or seldom/never true/certainly false}$ and $5 = \text{very often/very often true/certainly true}$). The BIS/BAS (Behavioral Activation Scale/Behavioral Inhibition Scale) uses a 4-point response option anchored by $1 = \text{very true for me}$ and $4 = \text{very false for me}$. W is the test statistic for comparing two reliability coefficients and can range from 0 to 1. Smaller numbers indicated larger and more statistically significant effects. It is impossible to compute W when one of the reliability coefficients is zero or negative. Thus, for the RFQ promotion scale, we set the alpha for the low education group at .01.

^aThe smaller sample for the NFC scale required that we compare the two Cronbach's alpha using the Fisher-Bonett test, which outputs a z -score instead of the W .