Does It Work for Everyone?
The Influence of Demographic Variables on Statistical Reliability

Andrew C. Pickett, PhD*
Danny Valdez, PhD
Adam E. Barry, PhD, FAAHB

Abstract

Recent developments have highlighted the importance of tailored health education efforts. However, little research has explored differential functioning of survey items for diverse populations. This work explores differences in statistical reliability for multiple scales across demographic groups. Understanding such differences is important in health research, given the rapid shifts occurring in global demographics. Study data were collected from eight years of the National College Health Assessment (*n* = 885,084), a large-scale annual survey of U.S. university students. Meta-analytic reliability generalization was used to compare reliability of two scale measures for multiple demographic groups. In nearly all cases, there were statistically significant differences in reliability across demographic groups. Researchers should consider relative functioning of any scale employed in their work. For certain demographic groups, various scales may not be sufficiently reliable. However, this may be obfuscated in larger samples, containing large numbers of individuals for whom the scale is sufficiently reliable. We suggest a thorough subsets analysis of data to ensure uniform functioning of items prior to use. Just as health interventions should be tailored to populations of interest, so too must research methods and tools.

*Corresponding author can be reached at: drew.pickett@usd.edu

The U.S. Census Bureau projects that by 2050, 54% of the U.S. population will identify with a racial group other than white (Colby & Ortman, 2014). Further, more than 20% of the U.S. population now speaks a language other than English at home (*Selected Social Characteristics in the United States*, 2017). Women, more than men, are projected to enter the labor workforce in the coming decades (Toossi & Morisi, 2017), and improved civil liberties are leading to increased civic engagement from under-represented populations, including sexual minorities and those with non-binary gender identities (Russell et al., 2010).

Consistent with these larger demographic shifts, U.S. college student populations are increasingly more diverse as well. Indeed, in the period from 1971 to 2018, the percentage of non-white students on college campuses rose from 10% to 47% (Osei, 2019). Similarly, according to the National Center for Educational Statistics, women now outnumber men among full-time college students at four-year institutions, representing 56% of the student population in 2016 (*National Postsecondary Student Aid Study*, 2016). Further, the National College Health Assessment recently reported that 20% of college students self-identified as a sexual minority (i.e., sexual identity other than ‘Heterosexual’) and 3.3% identified as gender non-binary (American College Health Association, 2018).

What does this increased diversity mean for those engaged in health behavior research, and should it influence the methodological choices they make? The purpose of this investigation was to examine whether health behavior measures function differently, depending on the characteristics of the respondent. Specifically, we sought to explore whether the reliability of health measures were different across respondent demographic classifications, for those who participated in eight years of the National College Health Assessment (NCHA; 2008 – 2015). Simply put, our research questions was, "Do commonly employed health behavior assessment items perform relatively better for certain groups?"

Score Reliability

To have confidence in their findings and associated recommendations, researchers must be cognizant of the psychometric properties of the measures they employ. One such property— reliability—describes the consistency of responses to items across individual participants and samples. Thus, a measure is said to be reliable if, and *only* if, it returns similar responses for individuals who are truly similar with respect to the phenomenon in question. Thompson explained measure reliability in the context of a bathroom scale (Thompson, 2002). He noted that if an individual were to step onto a bathroom scale multiple times in succession and receive wildly different estimates of their weight, the scale would be deemed unreliable. Conversely, if that same person stepped on and off the scale, each time receiving the same estimate of weight, this consistency would suggest the scale was reliable (or, at least, consistent in its level of inaccuracy). Noting the practical importance of reliability, Kerlinger and Lee asserted, "If one does not know the reliability […] of one's data, little faith can be put in the results obtained and the conclusions drawn from the results," (2000, p. 442). Thus, reliability is a vital component to the overall quality of conclusions health behavior scientists draw from their analyses, as well as resulting policy and practice recommendations.

When employed in a survey, there is an inherent assumption that scale items are sufficiently reliable for the particular researcher's purposes. Mathematically, all statistical analyses found within the general linear model (GLM) assume that data is perfectly reliable, with the exception of structural equation modeling (SEM) (Nimon, 2012). Of course, the assumption of perfect reliability is universally violated to some degree, given that observed data will always incur some amount of measurement error. Thus, it is left to individual researchers to determine the level (and acceptability) of unreliability found in their data. By conducting further statistical analyses on data, the researcher is operating from the assumption of acceptable reliability. To aid in this decision, several statisticians have suggested benchmarks of internal consistency (generally measured by Cronbach's Alpha), ranging from as low as .50 to as high as .90, in terms of "acceptability," (Henson, 2001; Nunnally, 1967, 1978, 1982).

By accepting a certain amount of unreliability across the entire sample, though, researchers may be obscuring systematic differences in reliability between subgroups. Unfortunately, relatively few scale measures are statistically validated for multiple populations. Indeed, most scale development and validation studies rely on some form of convenience sampling (e.g., psychology students who participate in studies for extra credit, snowball sampling), which have poor generalizability to begin with. As such, it could be argued that scale measures should be used only in the context of the population(s) from which they were developed, or be further validated for any new populations of interest to be studied. In other words, measures should be tested (and retested) for sub-populations to determine whether scales function similarly across groups.

Understanding differential functioning is important, as response patterns of one group may differ from those of other groups. Given their overall similarities, we expect individuals in homogenous groups to respond to scale measures in a similar pattern. However, we may also expect these groups to respond differently to each other. That is, a scale measure may be sufficiently reliable for use with one subset of the population, but not another. For example, one study noted that reliability of a measure of delinquent behaviors varied by the grade level of participants surveyed, such that the scale was more reliable for 8th and 10th year students than it was for 12th year students (Pickett et al., 2017). Intuitively, this makes sense, as younger children are unlikely to engage in delinquent behaviors (e.g., drug use, violence). Thus, the younger group is likely more homogenous with respect to these behaviors, as very few participants would have engaged in them at all. Conversely, within the older category, more individuals had likely engaged in a wider variety of delinquent behaviors—thereby increasing the overall variability of experiences and reducing the reliability of a scale measure for this group.

Overall, this study explores variations in statistical reliability, specifically seeking to demonstrate differential psychometric functioning of scale measures across demographic groups. We employ meta-analytic reliability generalization (RG) to test differences in reliability for various groups across multiple measures in a large, national (U.S.) sample of college students, over several years.

Method

Participants

Data for this study were drawn from the NCHA, which is a large national survey related to university students' health behaviors, collected twice annually. Across the NCHA's history, more than 1.7 million participants have completed surveys, drawn from more than 800 unique institutions (American College Health Association (ACHA), 2016). For the current investigation, we included all participants included within the NCHA survey dated from fall of 2008 to spring of 2015 (*n* = 885,084). This time period was chosen as major changes were made to the NCHA survey in both 2008 and 2015, including the re-wording of several items related to the current study. Thus, this date range represented the most recent time span (of sufficient length) in which direct comparison of scale reliability was possible, without confounding influences related to changes in phrasing. It is important to note that no exclusion criteria were applied to respondents, resulting in the data examined including undergraduate and graduate respondents, as well as full-time and part-time students attending a variety of colleges and universities (public and private) spread across the United States.

Measures

We explored the performance of items specifically examining a health risk factor (substance use), as well as a health protective factor (physical activity). Given college is typically characterized as a time of increased substance use behaviors, and a significant proportion of college students demonstrate behaviors that could be characterized as hazardous and at-risk, we specifically chose to examine alcohol, tobacco, and other drug use behaviors (National Institute on Drug Abuse (NIDA), 2018; Palmer et al., 2012). Additionally, despite the clear health benefit of physical activity, rates generally decline with age, often beginning in young adulthood (Caspersen et al., 2000; Dougall et al., 2011; Kilpatrick et al., 2005). This is important, as such declines are often permanent and return to highly active lifestyles is uncommon. These items were chosen due to the central purpose of the current study, namely, that we were interested in examining both a health behavior risk factor and protective behavior factor, to ensure factors did not share a single domain.

Frequency of Alcohol, Tobacco, and Other Drug (ATOD) Use. Participants' ATOD use was measured via five items (common across all iterations of NCHA), in which participants were asked to rate the frequency with which they had used several substances. Specifically, participates responded to the item stem: "In the last 30 days, on how many days did you use…" There were eight possible response options, including: "never used" (1); "have used, but not in last thirty days" (2); 1-2 days (3); 3-5 days (4); 6-9 days (5); 10-19 days (6); 20-29 days (7); and used daily (8). Substances assessed include alcohol, cigarettes, cigars, smokeless tobacco, and marijuana. Across all samples, these items demonstrated marginally acceptable internal consistency (α = .67).

Frequency of Physical Activity Behaviors. Participants' physical activity habits were measured using three items, rating the frequency with which they engaged in certain types of exercise. Participants were asked to record the number of days per week they engaged in: (1) moderate exercise of at least 30 minutes; (2) vigorous exercise for at least 20 minutes; and (3) exercise to strengthen muscle via weightlifting of 8-12 repetitions. Response options ranged from 0 to 7 days per week. Across the entire sample, these three items had acceptable overall internal consistency (α = .79).

Procedure

Given this investigation was a secondary data analysis and the authors did not directly collect data, this work was determined exempt from review by the Institutional Review Board (IRB) at the first author's institution. For each iteration of the study and demographic group tested, individual reliability estimates (i.e., Cronbach's α) were calculated using SPSS. These individual reliability estimates were then recorded to comprise data for further comparison. Comparisons across demographic variables for both scales were then made using analyses of variance, with Tukey's post hoc tests to further explore group differences in reliability in the case of a significant omnibus effect.

Results

This study sought to explore group differences in reliability based on various demographic factors. We present each factor, separated by demographic dimension, in turn, below.

ATOD Use

Gender. Gender was categorized using three self-identified categories (Female, Male, Transgender). There was a statistically significant difference in scale reliability across the gender categories [F(2,39) = 198.64, p < .001]. Post hoc testing revealed that each group was significantly different from each other, with the Trans category (α = .86) demonstrating the highest reliability, followed by Males (α = .69), and Females (α = .61). This dynamic can be seen in Figure 1 below, in which the reliability for transgender persons was uniformly higher than that of males, which itself was uniformly higher than that of females.

Race. Significant differences across racial categories were also observed [$F(5,78) = 8.23$, $p < .001$]. Multiple group differences can be seen in Table 1 below.

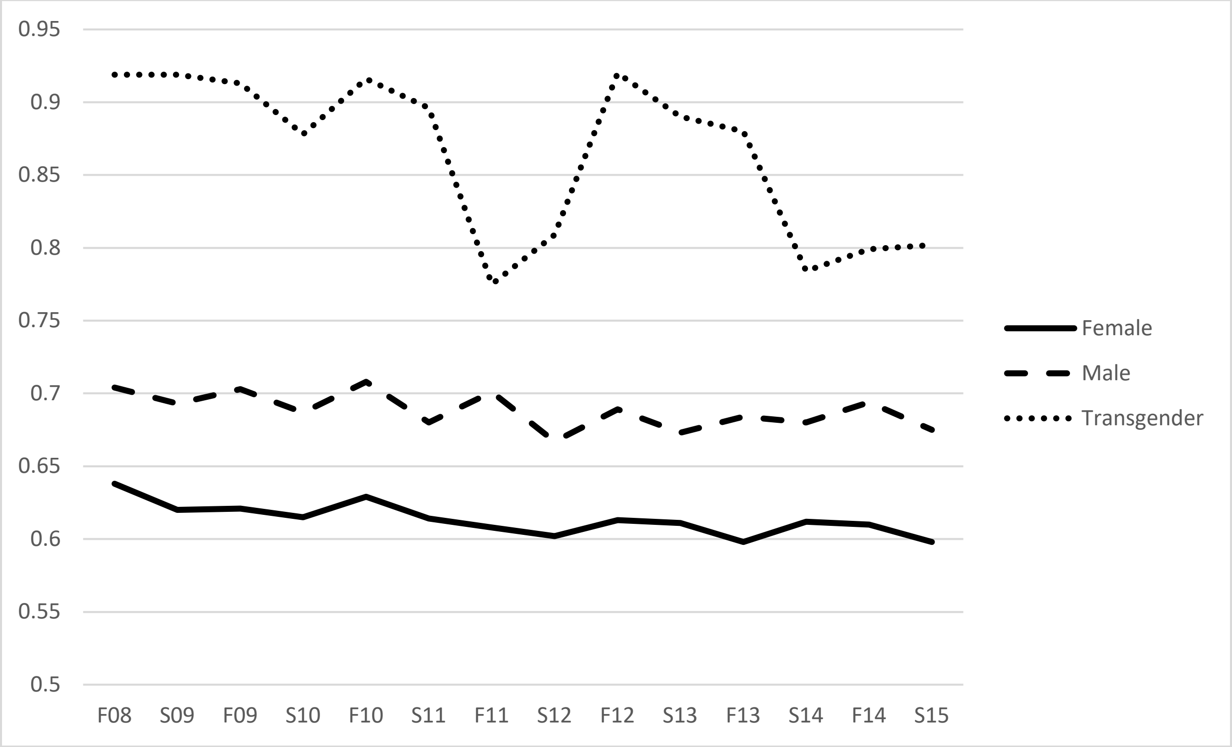

Figure 1. Plot of substance use scale reliability by gender.

Table 1

Post-hoc Analysis of Group Differences in Reliability of Substance Use by Race

Race	***n***	**1**	**2**	**3**
White	14	.65		
Black	14	.67		
Asian	14	.67	.67	
Latina/o	14	.68	.68	.68
Other	14		.70	.70
Native American	14			.70

Sexual Orientation. Participants were given four choices to best identify their sexual orientation (heterosexual, gay or lesbian, bisexual, and unsure). Again, significant differences in reliability by demographic group were observed [F(3,52) = 101.45, p < .001]. Post hoc analyses suggested three primary groupings of reliability, such that the highest internal consistency was seen among those identifying as unsure (α = .77). The next group included those identifying as bisexual and heterosexual (α = .66 and .65, respectively), which did not significantly differ from each other. Finally, those identifying as gay or lesbian had significantly lower reliability than all other groups (α = .62).

Year in school. The final grouping variable tested explored group differences by a participant's year in school. These were split into six categories, splitting first through fifth-year students individually and a final category for graduate students. The omnibus test suggested significant group differences in scale reliability [F(5,78) = 160.23, p < .001]. Post hoc testing suggested a number of significant differences, such that the scale became statistically significantly less reliable for each successive year in school (α ranging from .70 to .56), with fourth and fifth-year students being the only two groups not significantly different from each other.

Physical Activity

Gender. There were no observed statistically significant differences in reliability of the physical activity measure related to gender [F(2,39) = .56, p = .57]. The mean reliabilities for these groups were tightly clustered, ranging from .78 to .80. As seen in Figure 2, for this measure, there was no uniform pattern by which one group could be categorized as more reliable than the others. Specifically, the male and female categories were very similar across the years, with the reliability for trans persons showing greater variability, with scores both above and below their cisgender counterparts. Interestingly, this was the only variable, across both scales, for which no significant group differences were observed.

Race. While there were no observed gender effects related to the reliability of the physical activity measure, there were significant group differences related to race [F(5,78) = 11.78, p < .001]. Again, based on post hoc findings, three predominant subsets emerged between racial groups. Group differences and subsets can be seen in Table 2 below.

Sexual Orientation. As with race, there were significant differences in the reliability of the physical activity measure related to sexual orientation [F(3,52) = 10.46, p < .001]. For this measure, those identifying as heterosexual had the highest overall reliability (α = .79), which was significantly greater than all other groups. Gay and lesbian participants had the next highest internal consistency (α = .77), which was not significantly different from the unsure category (α = .76), but was significantly higher than the bisexual category (α = .74). The unsure and bisexual groups did not significantly differ from each other.

Year in school. Finally, we examined differences in reliability of the physical activity measure by year in school and found a significant omnibus effect [F(5,78) = 74.46, p < .001]. Similar to the substance use by year in school, reliability of the measure generally seemed to diminish as participants advanced in age. However, there were less distinct differences between categories for this scale.

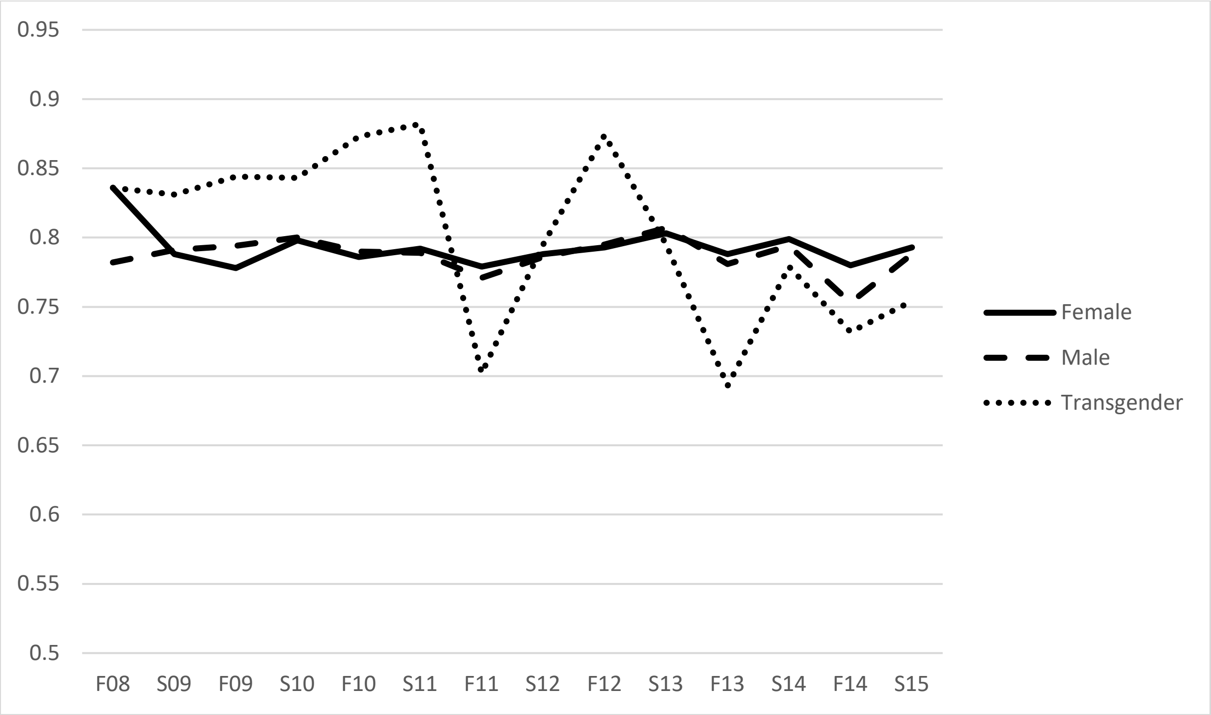

Figure 2. Plot of physical activity scale reliability by gender.

Table 2

Post-hoc Analysis of Group Differences in Reliability of Physical Activity by Race

Race	*n*	1	2	3
Asian	14	.78		
White	14	.78		
Other	14	.78		
Native American	14	.80	.80	
Latina/o	14		.81	.81
Black	14			.82

The three emergent subsets can be seen in Table 3.

Discussion

Nunally and Bernstein note, "it is meaningful to think of a test as having a number of different reliability coefficients, depending on which sources of measurement error are considered" (1994, p. 256). Simply put, the internal consistency of scores can vary due to an infinite number of possible confounding factors—many of which are random occurrences. However, in many cases, measurement error of items is systematic (i.e., *not* random). Here, we have

Table 3

Post-hoc Analysis of Group Differences in Reliability of Physical Activity by Year in School

Year	***n***	**1**	**2**	**3**
Graduate	14	.72		
5th Year	14		.78	
4th Year	14		.79	.79
3rd Year	14			.79
2nd Year	14			.80
1st Year	14			.80

sought to explore one source of measurement error in statistical reliability, sample diversity. In doing so, we found that in nearly every case, there were significant differences in reliability between groups. It would follow, then, that researchers using such measures in their research should mindfully consider the composition of their own samples prior to employing a scale measure. Further, researchers should consider the composition of samples used for validation studies, understanding that to some degree, deviation between their sample and those used during scale development will necessarily introduce new measurement error.

Further, researchers often use subsets of large datasets, such as the NCHA, to perform more specific analyses. For example, one may explore substance abuse among trans persons or physical activity of women. In such cases, it is important to understand the specific reliability of measures for that subset of participants. Using the traditional cut-off for acceptable scale reliability (i.e., α= .70), for example, certain scales analyzed here would be acceptable for some groups, but not for others. Understanding the differential functioning of such items for populations of interest is, therefore, vital to those analyzing subsets of large datasets.

With respect to frequency of ATOD use, we observed group differences related to gender, race, sexual orientation, and participants' year in school. That is, items functioned differently for groups across each of those dimensions. For example, as shown in Figure 1, this measure was uniformly most consistent for transgender persons, followed by cisgender men, and then cisgender women. This suggests that trans persons have more uniform substance use habits than their cisgender counterparts, or at least more uniformly respond to questions about frequency of use. This would be important for researchers interested in exploring substance use differences between trans and cisgender persons, as they should account for relative differences in reliability of the measure between groups.

However, one does not necessarily need to be testing differences between the groups in question for differential reliability to be of concern. If, for example, a study was instead concerned with exploring the role of substance use on students' grades, researchers would similarly benefit from understanding differential functioning of their items. Given the reliability of this measure for women always fell below the generally accepted .70 threshold, a sample with a high proportion of women may not have sufficient reliability to conduct further analyses. Even if that threshold was achieved, researchers should have less confidence in study findings, as its measures would still likely be less reliable than samples with higher representation of more internally consistent groups. This is not to say that understanding the effects of substance use on women's grades are unimportant, but rather the statistical findings of analyses employing this particular measure for women should be qualified due to its relatively poorer functioning for this group. Similar qualifications should be made with respect to all of the group differences observed here.

We also saw a number of differences among the physical activity variables. For these items, there were statistically significant group differences in reliability due to race, sexual orientation, and year in school. This is conceptually important due to the difference in domain (i.e., substance use vs. physical activity), as it suggests that group differences in reliability are not uniquely due to dynamics related to the content area of items. Instead, we argue that, for *any* scale measure, items will perform relatively better or worse for different demographic groups. Thus, it is important for researchers to understand such dynamics of any scale they choose to employ.

Traub has argued, "Reliability coefficients are almost always calculated from the measurements taken of samples or persons, not whole populations. […] Had a different sample of persons participated in the experiment, the reliability coefficient obtained would almost certainly have been a different number" (1994, p. 66). As such, various others have argued for the importance of reporting reliability estimates for individual samples, rather than estimates from test manuals, other samples, or validation studies (Pickett et al., 2017, 2019; Thompson, 2002). Here, however, we extend this call, suggesting authors further consider subsets analyses of reliability within their samples. That is, we encourage researchers to explore potentially differential function of scale measures related to demographics within their samples and to appropriately qualify findings. Group differences observed in the current study, across two scale measures, suggest that at least some variation in internal consistency of scale measures is systematic (i.e., not random) and related to study participants' various personal characteristics. Statistical conclusions drawn from such samples should be drawn with such limitations in mind.

Limitations and future research. The current study, despite using data from a large national sample, had limitations that should be acknowledged. First, institutions self-select to participate in the NCHA process. Thus, though it is appropriate to characterize this sample as national in scope, it would be inappropriate to identify it as nationally representative. While Hispanic serving institutions, and historically black colleges and universities participate, the majority of the sample is comprised of primarily white institutions. Second, we were limited in our choice of variables from which to draw. With regard to the behaviors examined (frequency of ATOD use and physical activity), the items employed may not have been specifically developed to constitute a "scale." In other words, these items—though employed by the NCHA for many years—may not have been developed as a comprehensive set of items intended to capture the universe of content associated with the behaviors. Nevertheless, we believe these items are worth exploration given (a) they highlight important health factors among the sample of interest, and (b) they have been employed in long-term surveillance efforts over several decades. Further, while we could not measure their effects due to the items available, we suspect that other demographic variables would also influence reliability of scale measures (e.g., age, socio-economic status, language). This is consistent with others, who have noted a myriad of external factors that may influence overall reliability of a measure (Ursachi et al., 2015). This suggests a large amount of future work—particularly with respect to commonly employed scales in health behavior research. In particular, we argue that it is vitally important for subsets analyses to be conducted to better understand the relative reliability of the measure for various groups—and accordingly, the relative appropriateness of the use of these scale measures for such groups.

Implications for Health Behavior Theory

Researchers in health education regularly employ scale measures to quantify complex, underlying latent variables. By definition, the findings of such studies are limited by the global reliability of the measures used. However, we contend that researchers should be concerned with the differential reliability of scale measures between demographic groups. Our results suggest that commonly employed health behavior items may, for example, be sufficiently reliable for one demographic group but not another. Thus, researchers seeking to employ such scales should understand the relative functionality of these measures for various subgroups present in their samples. Specifically, we recommend researchers use subsets analyses to ensure the reliability of scale measures for relevant demographic groups in their samples, and appropriately qualify findings. Given the fundamental importance of reliability in the use of scale measures, it should be similarly important to test for systematic measurement error related to demographics of a sample. Further, editors and reviewers should be aware of the potential for such dynamics in research submitted for publication and, when relevant, seek additional information from authors regarding any potentially differently functioning items. These general safeguards, focused on increased understanding and reporting of statistical reliability, are important for health behavior research.

Discussion Questions

1. Our findings suggest scale measures may function differently for various demographic subgroups. What new editorial practices can (or should) be implemented to ensure that measures used are sufficiently reliable for all relevant subgroups within a sample?
2. We suggest commonly reported estimates of statistical reliability may obfuscate systematic error in measures. What other systematic biases may go unnoticed, despite authors meeting common research reporting standards?

Acknowledgments

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. As such, the authors report no conflicts of interest, financial or otherwise.

References

American College Health Association (ACHA). (2016, March 14). *ACHA-NCHA Data*. American College Health Association. http://www.acha-ncha.org/

American College Health Association (ACHA). (2018). *National College Health Assessment (NCHA)*. https://www.acha.org/NCHA/Home/NCHA/NCHA_Home.aspx?hkey=f8184410-19fa-4ba6-b791-43a79cef2de0

Caspersen, C. J., Pereira, M. A., & Curran, K. M. (2000). Changes in physical activity patterns in the United States, by sex and cross-sectional age. *Medicine & Science in Sports & Exercise*, *32*(9), 1601–1609. https://doi.org/10.1097/00005768-200009000-00013

Colby, S. L., & Ortman, J. M. (2014). *Projections of the size and composition of the U.S. population: 2014 to 2060* (No. P25-1143; Current Population Reports, p. 13). Washington, DC: United States Census Bureau. https://www.census.gov/library/publications/2015/demo/p25-1143.html

Dougall, A. L., Swanson, J. N., Grimm, J. R., Jenney, C. T., & Frame, M. C. (2011). Tempering the decline in college student physical activity using informational interventions: Moderating effects of stress and stage of change. *Journal of Applied Biobehavioral Research*, *16*(1), 16–41. https://doi.org/10.1111/j.1751-9861.2011.00064.x

Henson, R. K. (2001). Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. (Methods, plainly speaking). *Measurement and Evaluation in Counseling and Development*, *34*(3), 177–189. https://doi.org/10.1080/07481756.2002.12069034

Kerlinger, F. N., & Lee, H. B. (2000). *Foundations of behavioral research* (4th ed.). Belmont, CA: Wadsworth Publishing.

Kilpatrick, M., Hebert, E., & Bartholomew, J. (2005). College students' motivation for physical activity: Differentiating men's and women's motives for sport participation and exercise. *Journal of American College Health*, *54*(2), 87–94. https://doi.org/10.3200/JACH.54.2.87-94

National Postsecondary Student Aid Study. (2016). National Center for Education Statistics. (2018, February 15). *College-Age & Young Adults*. https://nces.ed.gov/surveys/npsas/

Nimon, K. F. (2012). Statistical assumptions of substantive analyses across the general linear model: A mini-review. *Frontiers in Psychology*, *3*, 322. https://doi.org/10.3389/fpsyg.2012.00322

Nunnally, J. C. (1967). *Psychometric theory*. New York, NY: McGraw-Hill.

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

Nunnally, J. C. (1982). Reliability of measurement. In H. E. Mitzel (Ed.), *Encyclopedia of educational research* (pp. 1589–1601). New York, NY: Free Press.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd edition). New York, NY: McGraw-Hill.

Osei, Z. (2019). Low-income and minority students are growing share of enrollments, and 2 other takeaways from new study. *Chronicle of Higher Education*. https://www.chronicle.com/article/Low-IncomeMinority/246346

Palmer, R. S., McMahon, T. J., Moreggi, D. I., Rounsaville, B. J., & Ball, S. A. (2012). College student drug use: Patterns, concerns, consequences, and interest in intervention. *Journal of College Student Development*, *53*(1), 124-132. https://doi.org/10.1353/csd.2012.0014

Pickett, A. C., Valdez, D., & Barry, A. E. (2017). Psychometrics matter in health behavior: A long-term reliability generalization study. *American Journal of Health Behavior*, *41*(5), 544–552. https://doi.org/10.5993/AJHB.41.5.3

Pickett, A. C., Valdez, D., & Barry, A. E. (2019). Measurement implications associated with refinement of sexual and gender identity survey items: A case study of the National College Health Assessment. *Journal of American College Health*, 1–7. https://doi.org/10.1080/07448481.2019.1598421

Russell, S. T., Toomey, R. B., Crockett, J. L., & Laub, C. (2010). LGBT politics, youth activism, and civic engagement. In L. R. Sherrod, J. Torney-Purta, & C. A. Flanagan (Eds.), *Handbook of research on civic engagement in youth* (pp. 471–496). Hoboken, NJ: John Wiley and Sons.

Thompson, B. (2002). *Score reliability: Contemporary thinking on reliability issues*. Thousand Oaks, CA: SAGE Publications.

Selected social characteristics in the United States (American Community Survey). (2017). United States Census Bureau. https://data.census.gov/cedsci/

Toossi, M., & Morisi, T. L. (2017). *Women in the workforce before, during, and after the Great Recession* (Spotlight on Statistics, pp. 1–21). Washington, DC: United States Bureau of Labor Statistics.

Traub, R. E. (1994). *Reliability for the social sciences: Theory and applications*. Thousand Oaks, CA: SAGE Publications.

Ursachi, G., Horodnic, I. A., & Zait, A. (2015). How reliable are measurement scales? External factors with indirect influence on reliability estimators. *Procedia Economics and Finance*, *20*, 679–686. https://doi.org/10.1016/S2212-5671(15)00123-9