In Reply: Drs Patterson and Yealy raise questions about the psychometric properties of the assessments used in our study. It is indeed important to have appropriately valid and reliable measures of the constructs under study for any research. The literature regarding the issues of comparative psychometric performance of single- vs multiple-item assessments indicates that there are situations for which either or both are appropriate.[1,2]

The single-item measure of fatigue used in our study is derived from the well-known Brief Fatigue Inventory,[3] which has been used extensively as a primary end point in clinical studies. This single item has recently been found to be prognostic for survival among cancer patients and is superior psychometrically to longer multi-item measures in detecting individual change over time.[4] As a result of its demonstrated psychometric and prognostic capability, this fatigue item is now a standard measure of fatigue in North Central Cancer Treatment Group clinical trials.

The Epworth Sleepiness Scale has extensive psychometric documentation supporting its validity as a measure of excessive daytime sleepiness. The Epworth Sleepiness Scale has been reported to be the most discriminating test of daytime sleepiness[5] and a stable measure of sleep quality and sleepiness.[6]

Our study found that fatigue, sleepiness, and distress in resident physician respondents are each associated with the likelihood of subsequent self-reported major medical errors. We acknowledge that more nuanced insights into complex constructs such as fatigue and sleepiness may be possible using alternative survey instruments. However, our results are based on well-validated survey measures.

Jeff A. Sloan, PhD
Department of Health Sciences Research
Tait D. Shanafelt, MD
Colin P. West, MD, PhD
west.colin@mayo.edu
Department of Internal Medicine
Mayo Clinic
Rochester, Minnesota

1. Sloan JA, Aaronson N, Cappelleri JC, Fairclough DL, Varricchio C; Clinical Significance Consensus Meeting Group. Assessing the clinical significance of single items relative to summated scores. *Mayo Clin Proc*. 2002;77(5):479-487.
2. Huschka MM, Mandrekar SJ, Schaefer PL, Jett JR, Sloan JA. A pooled analysis of quality of life measures and adverse events data in North Central Cancer Treatment Group lung cancer clinical trials. *Cancer*. 2007;109(4):787-795.
3. Mendoza TR, Wang XS, Cleeland CS, et al. The rapid assessment of fatigue severity in cancer patients: use of the Brief Fatigue Inventory. *Cancer*. 1999; 85(5):1186-1196.
4. Sloan JA, Liu H, Sargent DJ, et al. A patient-level pooled analysis of the prognostic significance of baseline fatigue for overall survival among 3,915 patients participating in 43 North Central Cancer Treatment Group and Mayo Clinic Cancer Center oncology clinical trials [abstract 9599]. *J Clin Oncol*. 2009;27(15s):9599.
5. Johns MW. Sensitivity and specificity of the multiple sleep latency test (MSLT), the maintenance of wakefulness test and the Epworth sleepiness scale: failure of the MSLT as a gold standard. *J Sleep Res*. 2000;9(1):5-11.
6. Knutson KL, Rathouz PJ, Yan LL, Liu K, Lauderdale DS. Stability of Pittsburgh Sleep Quality Index and the Epworth Sleepiness Questionnaires over 1 year in early middle-aged adults: the CARDIA study. *Sleep*. 2006;29(11):1503-1506.

Mindful Communication to Address Burnout, Empathy, and Attitudes

To the Editor: We have 2 concerns about the outcome measures and the statistical methods used in the study of the association of mindful communication with burnout, empathy, and attitudes among primary care physicians by Dr Krasner and colleagues.[1] First, the outcome measures may have been influenced by social desirability responding, the tendency for participants to present a favorable image of themselves. Participants completed 5 sets of Likert surveys covering socially sensitive topics, which might have been contaminated by social desirability responding. This may happen, for example, if persons with higher levels of social desirability responding substitute a socially desirable response for the true response; in that case, any variation in responses would reflect variations in understanding the social norms rather than variations in personality. Social desirability responding can suppress, obscure, or produce artificial relationships among variables, resulting in confounding.[2] The study design does not appear to have controlled for or distinguished social desirability responding. It may be possible to explore the relationship between 2 variables of interest while statistically controlling for social desirability responding using partial correlations.[3]

Second, although linear mixed-effects models were used in the statistical analysis, the organizational effects among individuals might not have been considered. Linear mixed-effects models help account for the nesting of repeated measures within individuals.[4] However, several of the 70 persons enrolled might have been from the same health care organization, which would challenge the utility of linear mixed-effects models. If the characteristics of the enrolled individuals were not independent before and during the educational program, the group effect should also be disaggregated. When the authors stated that it was possible that the observed changes resulted primarily from their participants' spending time together with their colleagues, they were not referring to colleagues who were not enrolled. Hence, we believe that modeling both within- and between-individual effects can be used to mine the data further.

Peng Wang, PhD
pengsdnu@yahoo.com.cn
Fengqiang Gao, PhD
College of Psychology
Shandong Normal University
Ji Nan, China

1. Krasner MS, Epstein RM, Beckman H, et al. Association of an educational program in mindful communication with burnout, empathy, and attitudes among primary care physicians. *JAMA*. 2009;302(12):1284-1293.
2. King M, Bruner G. Social desirability bias: a neglected aspect of validity testing. *Psychol Mark*. 2000;17(2):79-103.
3. Puhl R, Schwartz M, Brownell K. Impact of perceived consensus on stereotypes about obese people: a new approach for reducing bias. *Health Psychol*. 2005; 24(5):517-525.

4. Burton P, Gurrin L, Sly P. Extending the simple linear regression model to account for correlated responses: an introduction to generalized estimating equations and multi-level mixed modelling. *Stat Med*. 1998;17(11):1261-1291.

In Reply: Drs Wang and Gao assert that social desirability is a response bias that can confound relationships among the variables of interest. This claim is nonspecific and belied by the evidence. First, a wide body of psychometric data suggests that the putative social desirability response bias is not a response bias at all, but a misnomer for a personality trait related to conventionality.[1] Second, even if social desirability resulted in a response bias, there is no psychometric consensus on how it should be measured.[2] Third, even if its measurement were attempted, including the results of this measurement in models may reduce rather than enhance the validity of associations.[2,3] For instance, in the legally perilous and financially high-stakes world of personnel selection, meta-analytic evidence weighs heavily against the claims that social desirability is a response bias and confounds associations of interest; efforts to measure and adjust for it are discouraged.[3] Fourth, even if a social desirability bias exists, measurement is endeavored, and adjustment attempted, social desirability responsiveness could bias treatment effect estimates only if it were associated with both the outcome and treatment.

Even if all of the preceding conditions prevailed, data on the effects of mindfulness suggest that this putative bias would likely work in the opposite direction of the study's findings. The magnitude of the treatment effect would be underestimated because mindfulness treatments reduce, rather than increase, the reactivity to social norms and close-mindedness characteristic of the social desirability bias.[4] Therefore, social desirability responding would operate prior to treatment, inflating pretreatment reports of adjustment. Because treatment reduces attributes giving rise to this putative response bias, posttreatment reports of adjustment would not benefit from the same inflation as pretreatment reports, thereby spuriously reducing the difference between pretreatment and posttreatment scores.

Wang and Gao also raise concerns regarding the possible organizational effects among individuals, suggesting that the group effect should also be disaggregated and that modeling of both within- and between-individual effects can be used. First, we note that there were no other group effects to be disaggregated. Second, the linear mixed models used already model between- and within-individual effects. Third, even if clustering on a variable is observed, specification of a random effect is only appropriate if the clustering can be traced to the data generating process.[5] A nonzero intraclass correlation is necessary but not sufficient evidence of such a data generating process. The sufficient condition is that measurements are random variables arising from a probability space in which their expected probabilities, conditional on the grouping factor, are different from their unconditional expected probabilities.[5] Our study did not exhibit such a data generating process. Mining the data for additional random effects would therefore lead to model misspecification and compromised statistical inference.

Benjamin P. Chapman, PhD
Department of Psychiatry
Michael S. Krasner, MD
michael_krasner@urmc.rochester.edu
Department of Internal Medicine
Ronald M. Epstein, MD
Department of Family Medicine
University of Rochester School of Medicine and Dentistry
Rochester, New York

1. McCrae RR, Costa PT. Social desirability scales: more substance than style. *J Consult Clin Psychol*. 1983;51(6):882-888.
2. Barger SD. The Marlowe-Crowne affair: Short forms, psychometric structure, and social desirability. *J Pers Assess*. 2002;79(2):286-305.
3. Ones DS, Viswesvaran C, Reiss AD. Role of social desirability in personality testing for personnel selection: the red herring. *J Appl Psychol*. 1996;81(6):660-679.
4. Brown KW, Ryan RM, Creswell JD. Mindfulness: theoretical foundations and evidence for its salutary effects. *Psychol Inq*. 2007;18(4):211-237.
5. Spanos A. *Probability Theory and Statistical Inference: Econometric Modeling With Observational Data*. Cambridge, UK: Cambridge University Press; 1999:145-173.

Tools to Assess Clinical Skills of Medical Trainees

To the Editor: Dr Kogan and colleagues[1] reported the results of a systematic review of observation tools used to assess clinical skills of medical trainees, summarizing the evidence of the tools' validity and outcomes. The authors stated in their methods that every article was independently abstracted by 2 authors and that differences in data abstraction were resolved through consensus adjudication.

Regarding the differences in data abstraction, I would like to know what was the proportion of differences that required adjudication and if those differences were related to specific items of the abstraction form. In their review, the authors did not provide quantitative measures of the inter-rater reliability; reporting κ coefficients for the main review items would have been informative. I would also like to know what the authors meant by consensus adjudication (methods and number of persons involved).

In addition, the authors used a modification of the Best Evidence in Medical Education abstraction form to collect their data. This modified abstraction form could have been tested on a randomly selected sample of articles to check the adequacy of the instrument and to train the raters.

Isabelle Pitrou, MD, MSc
isabellepitrou@yahoo.fr
René Descartes University
Paris, France

1. Kogan JR, Holmboe ES, Hauer KE. Tools for direct observation and assessment of clinical skills of medical trainees: a systematic review. *JAMA*. 2009;302(12):1316-1326.