Sources of Comparability Between Probability Sample Estimates and Nonprobability Web Sample Estimates

William Riley[1], Ron D. Hays[2], Robert M. Kaplan[1], David Cella[3],

[1]National Institutes of Health, Bethesda, MD; [2]University of California Los Angeles, Los Angeles, CA; [3]Northwestern University, Chicago, IL

Proceedings of the 2013 Federal Committee on Statistical Methodology (FCSM) Research Conference

Introduction

Unbiased population based estimates of health status are required for effective planning and population monitoring. The strengths and weaknesses of nationally representative probability samples versus non-probability Internet panels is increasingly debated (Baker et al., 2013). Probability samples are assumed the gold standard of surveillance research. It is difficult, however, to execute probability sampling in dual-frame random digit dialing surveys (RDD) and often costly to implement. The ubiquity of caller ID and the shift from landlines to cell phones as the primary telecommunications device substantially limits the use RDD as a probability sampling methodology (Voigt, Schwartz, Doody, Lee, & Li, 2011). Response rates among probability samples are quite low, leading some to argue that there is little practical difference between opting out of a probability sample versus opting in to a non-probability sample (Gotway Crawford, 2013; Rivers, 2013).

Surveys using non-probability Internet panel samples are substantially less expensive to conduct than probability samples, but there are legitimate concerns about sampling bias. Internet panel technologies and sampling methods have advanced, and some Internet panels approximate samples based on probability sampling (e.g. www.knowledgenetworks.com). Comparisons of responses from probability and non-probability samples can provide evidence about the extent of equivalence between these two sampling approaches.

In this study we compare the ten global items from the Patient-Reported Outcomes Measurement Information System (PROMIS) across a range of probability and non-probability surveys including the National Health Interview Survey (NHIS), HealthStyles, and pilot data from the Division of Behavioral Surveillance (DBS), Population Health Surveillance and Informatics Program Office (PHSIPO), Centers

for Disease Control and Prevention (CDC). The DBS Internet Opt-in panel pilots were administered by YouGov/Polimetrix to samples constructed to be representative of the national population.

Overview of PROMIS Global Items

PROMIS is a National Institutes of Health (NIH) initiative to utilize Item Response Theory (IRT) and Computer Adaptive Testing to develop and automate the administration of efficient, precise, and valid item banks measuring patient-reported clinical outcomes (e.g. pain, fatigue, physical function, depression) (Cella, et al., 2010). The PROMIS measures include 10 items that assess general perceptions of health (Hays, Bjorner, Revicki, Spritzer, & Cella, 2009). These Global Health items consist of 5 that assess general health items (e.g., In general, would you say your health is: excellent, very good, good, fair, or poor), and 5 derived from the core domains of the initial PROMIS item banks (e.g. physical function, pain, fatigue, emotional distress, social activities).

These items were tested using YouGov/Polimetrix, a non-probability Internet panel, augmented by clinical samples obtained from the PROMIS network (N = 21,333). Participants ranged in age from 18-100 years (mean age = 53) and 52% were female. Hispanics and African Americans each made up 9% of the sample. Fifty-nine percent were married. Nineteen percent had 12 or less years of education (Table 1). This sample was used to test and calibrate the PROMIS Global items as part of the first set of PROMIS item banks.

Exploratory and confirmatory factor analyses of the PROMIS Global items revealed two factors, physical health and mental health. Coefficient alphas were 0.81 and 0.86 for the 4-item physical health and 4-item mental health scales, respectively. The PROMIS Global Physical and Mental scores were scored using an IRT graded response model and scaled using a T-score metric (mean of 50, SD of 10). Because the PROMIS item testing and calibration sample was predominantly a non-probability Internet sample, a normative raking procedure was performed for norming and setting the T-scores (Liu, Cella, Gershon, Shen, Morales, Riley, & Hays, 2010). The Global Physical Health scale correlated highly with the PROMIS Pain Impact (r = -.75), Fatigue (r = -.73), Physical Function (r = -.71), and Pain Behavior (r = -

.67) item banks. The Global Mental Health scale correlated highly with the PROMIS Depression (r = -.71), Anxiety (r = -.65), and Satisfaction with Social Activities (r = -.60) item banks.

PROMIS Global Health Score Estimates Using Different Sampling Methodologies

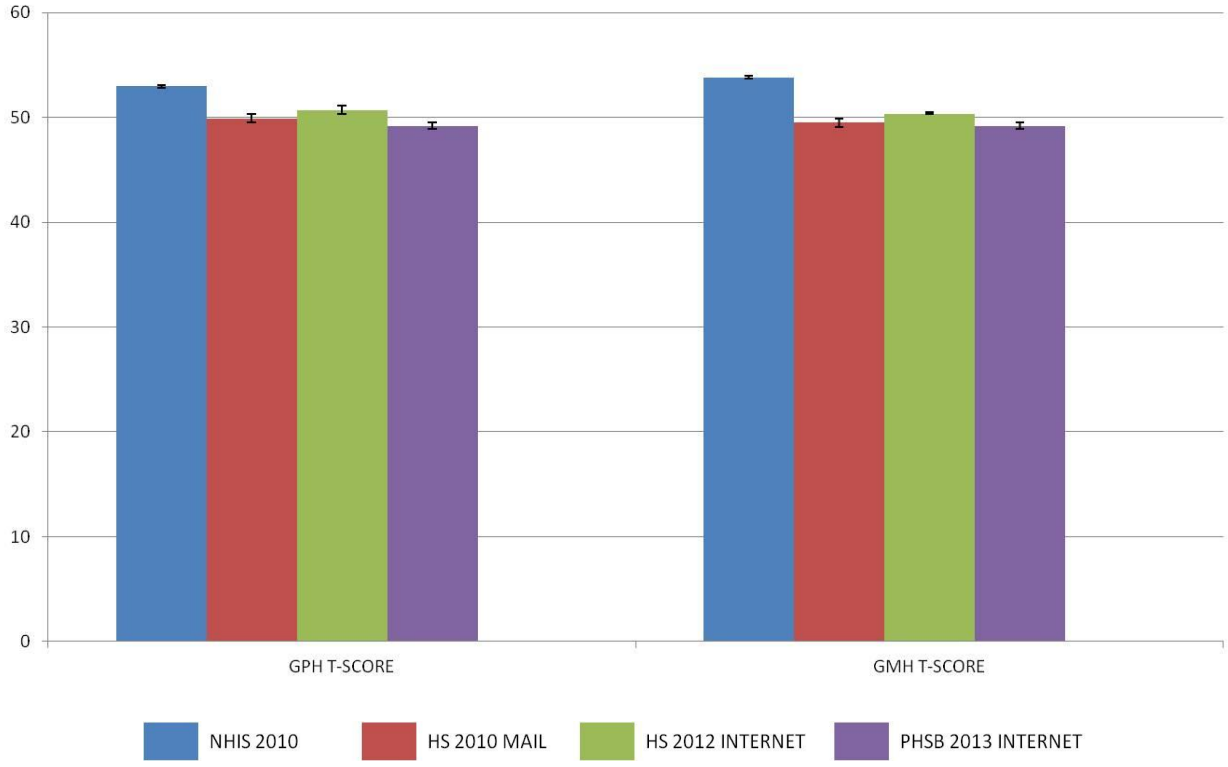
The PROMIS Global Health items were included in a number of national surveys including the National Health Interview Survey (NHIS), a nationally representative probability survey of households in the United States. (see Table 1). The items were also included in two cohorts of the HealthStyles survey, one in 2010 via a non-probability mail sample, and again in 2012 via a probability Internet panel. Finally, the PROMIS Global Health items were piloted in four U.S. states and metropolitan statistical areas (MSAs) as well as nationally by the DBS/PHSIPO/CDC, using a non-probability sample constructed to be representative of each of these respective geographies (indicated in Table 1 as PHSB 2013).

Table 1: Population Characteristics

	PHSB 2013 (n=3,500)			HealthStyles 2012 (n=3,503)			HealthStyles 2010 (n=4,184)			NHIS 2010 (n=27,157)		
	Unweighted		Weighted	Unweighted		Weighted	Unweighted		Weighted	Unweighted		Weighted
	n	%	%	n	%	%	n	%	%	n	%	%
Race or Ethnicity												
Non-Hispanic White	2,635	75.3	67.4	2,641	75.4	67.0	2,842	68.0	69.0	15,510	57.2	68.4
Non-Hispanic Black	326	9.3	11.4	334	9.5	11.5	477	11.0	12.0	**4,394**	**16.2**	11.65
Hispanic	311	8.9	14.1	116	9.5	14.4	495	12.0	14.0	**5,054**	**18.6**	13.7
Other	228	6.5	7.2	412	5.6	7.1	370	9.0	6.0	2,171	8.0	6.3
Gender												
Female	1,968	56.2	52.0	1,770	50.5	51.7	2,181	52.0	52.0	15,171	**55.9**	51.7
Male	1,532	43.8	48.0	1,733	49.5	48.3	2,003	48.0	48.0	11,986	44.1	48.3
Age of Respondent												
18-24	283	8.1	13.1	317	9.0	12.6	60	**1.0**	13.0	2,801	10.3	12.8
25-34	565	16.1	15.7	418	11.9	17.2	414	10.0	18.0	4,974	18.3	17.9
35-44	786	22.5	18.9	518	14.8	17.2	707	17.0	18.0	4,805	17.7	17.4
45-54	533	15.2	16.8	718	20.5	18.9	1,269	**30.0**	20.0	4,855	17.9	19.4
55-64	751	21.5	18.2	706	20.2	16.2	806	19.0	15.0	4,272	15.7	15.6
65+	582	16.6	17.3	826	23.6	17.9	928	22.0	17.0	5,450	20.1	16.9

We compared PROMIS Global Health scores obtained from these different sampling approaches and modes of administration. As shown in Figure 1, the physical health and mental health T-score means and standard deviations were comparable across these four surveys, with the exception of the NHIS, which had approximately a .3 SD higher mean score than the other surveys.

Figure 1: T-Scores for PROMIS Global Physical Health (GPH) and Mental Health (GMH) by Survey



Figures 2 and 3 summarize responses to each of the PROMIS Global Physical Health and Mental Health items by survey. Responses to each of these items were generally comparable across surveys, with the notable exception of higher endorsement of the healthiest response option among those in the NHIS survey.

Figure 2: PROMIS Global Physical Health Item Responses by Survey

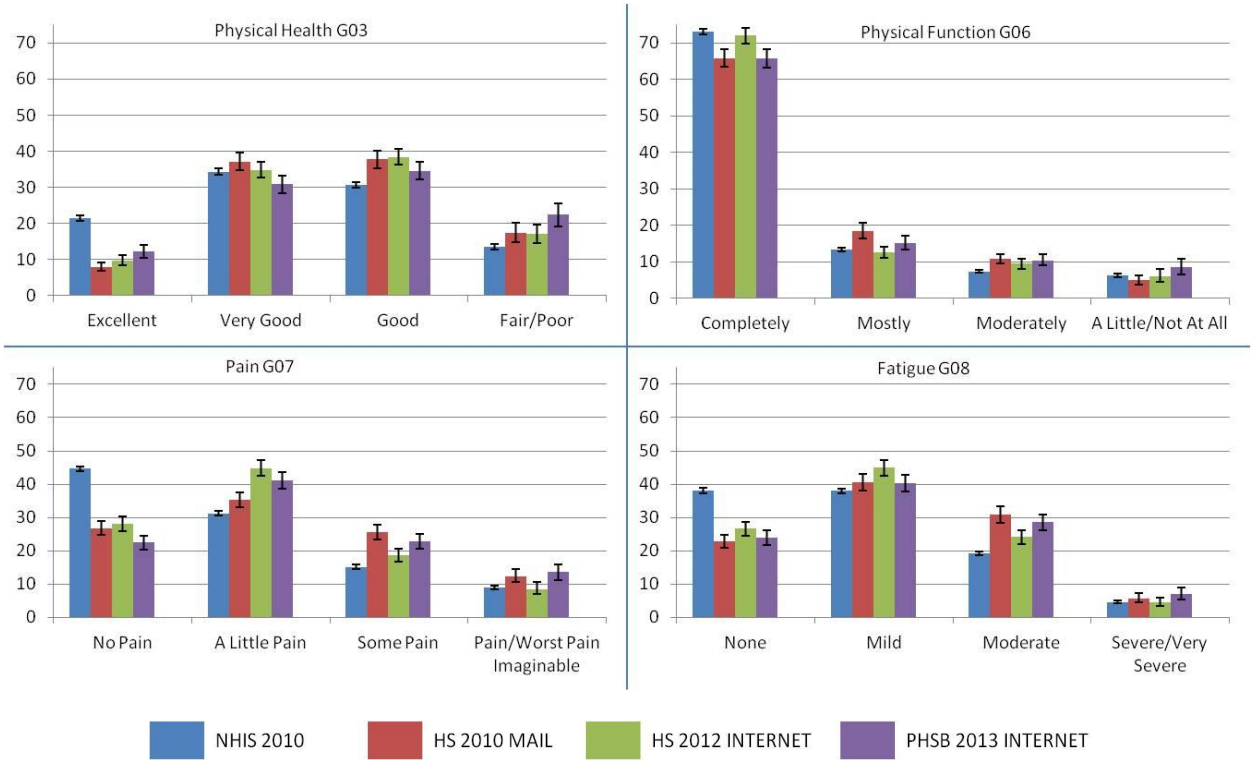
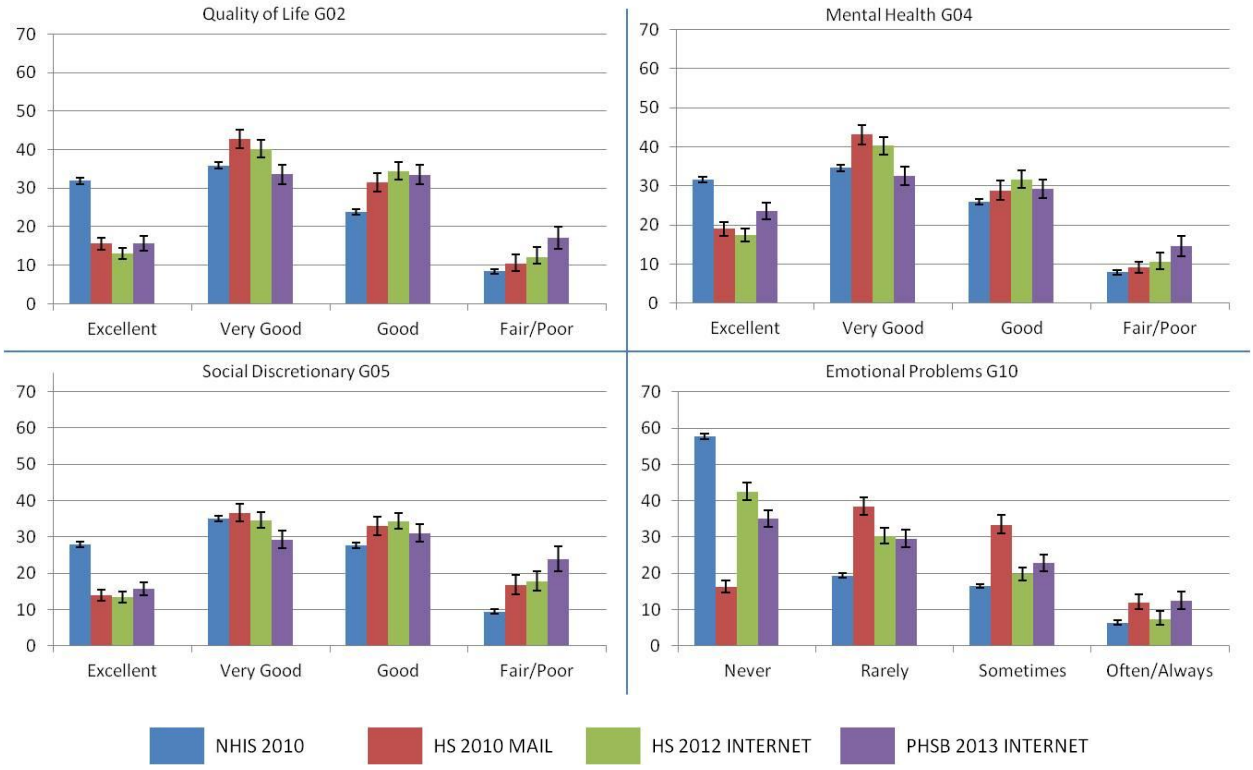


Figure 3: PROMIS Global Mental Health Item Responses by Survey



Discussion and Conclusions

The results of the comparison of the PROMIS Global Health items across four surveys reveals comparable levels of physical and mental health despite differences in survey sampling (probability or non-probability) and mode of administration (interview, mail, Internet). The notable exception was the NHIS, which produced higher mean scores on the PROMIS Global Physical and Mental Health scales. Examination of the response option patterns across individual items revealed a tendency for respondents in the NHIS to use the healthiest response option more frequently than respondents in the other surveys. It is not possible from these data to determine the source of this difference, but the most likely possibility is the interview mode of administration. Other studies have shown a 0.2 to 0.5 SD increase in health-related quality of life responses from in-person administration vs. mail or internet responses (Hays, et al., 2009). While we cannot rule out an effect of NHIS sampling methodology for this difference, it is important to note that there were minimal differences on the PROMIS Global Health item responses and scale scores between the other three surveys. These other three surveys included both probability and non-probability sampling, but were similar in their non-personal mode of administration (e.g. mail or internet).

Probability samples will remain a critical tool in surveillance research, but the combination of efficient data collection, ability to recruit targeted samples, and ease of replication has increased the value of non-probability samples. Via the raking procedure used for PROMIS or other weighting adjustments for non-probability samples (e.g. cell weighting, propensity score weighting) it may be possible to provide reasonable estimates of national populations from non-probability samples. Precision of estimates can be calculated from non-probability samples using Bayesian Credibility Intervals (Roshwalb, El-Dash, & Young, 2012). Non-probability samples have clear weaknesses including selection biases and the considerable variability in the procedures of various Internet polling panels, but the results of this study suggest that there is likely more variation between samples due to mode of administration than to the sampling methodology.

References

Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J.A., Gile, K. J., & and Tourangeau., R. (2013). Summary Report of the AAPOR Task Force on Non-probability Sampling *Journal of Survey Statistics and Methodology,1,* 90-143.

Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., Amtmann, D., Bode, R., Buysse, D., Choi, S., Cook, K., Devellis, R., DeWalt, D., Fries, J. F., Gerson, R., Hahn, E. A., Lai, J. S., Pilkonis, P., Revicki, D., Rose, M., Weinfurt, K., Hays, R., PROMIS Cooperative Group (2010). The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005-2008. *Journal of Clinical Epidemiology, 63,* 1179-1194.

Gotway Crawford, C. A. (2013). Comment. *Journal of Survey Statistics and Methodology, 1,* 118-124.

Hays, R. D., Bjorner, J.B., Revicki, D. A., Spritzer, K. L., & Cella, D. (2009). Development of physical and mental health summary scores from the Patient-Reported Outcomes Measurement Information System (PROMIS) global items. *Quality of Life Research, 18,* 873-880.

Hays, R. D., Kim, S., Spritzer, K. L., Kaplan, R. M., Tally, S., Feeny, D., Liu, H., & Fryback, D. G. (2009). Effect of mode and order of administration on generic Health-Related Quality of Life scores. *Values in Health, 12,* 1035-1039

Liu, H., Cella, D., Gershon, R., Shen, J., Morales, L. S., Riley, W., & Hays, R. D. (2010). Representativeness of the Patient-Reported Outcomes Measurement Information System Internet panel. *Journal of Clinical Epidemiology, 63*, 1169-1178.

Rivers, D. (2013). Comment. *Journal of Survey Statistics and Methodology, 1,* 111-117.

Roshwalb A., El-Dash, N., & Young, C. (2012). Towards the use of Bayesian credibility intervals in online survey results. NY: Ipsos Public Affairs, http://www.ipsos-na.com/dl/pdf/knowledge-ideas/public-affairs/IpsosPA_POV_BayesianCredibilityIntervals.pdf (Accessed February 19, 2014).

Voigt, L. F., Schwartz, S. M., Doody, D. R., Lee, S. C., & Li, C. L. (2011). Feasibility of including cellular telephone numbers in random digit dialing for epidemiologic case-control studies. *American Journal of Epidemiology, 173,* 118-126.

Acknowledgements: We thank Satvinder Dhingra, Catherine Okoro, and William Thompson from the Centers for Disease Control for their critical work on this project.