



# Psychometric Comparability of Single Item and Grid Form Administration of the SF-36v2 Health Survey

Kevin J. Smith, PhD; Mark Kosinski, MA • QualityMetric Incorporated, Lincoln, RI, USA



## OBJECTIVE

Recently, the search for more reliable and affordable data collection methods has been facilitated by the rapid growth of internet connectedness and availability of inexpensive personal computing technology. Recent guidance indicating the acceptability of electronic data collection by governing bodies[1,2] has spurred growing enthusiasm for the use of electronic patient reported outcomes (e-PRO) methods. While existing research supports the score equivalence of paper and electronic administration modes[3], it is well accepted that changes to the formatting and display characteristics of survey items could result in changes to the reliability and validity of survey results. Despite this, evidence concerning the impact of changes in item format, including those typically made to accommodate e-PRO, is limited even for the most widely used patient reported outcomes (PRO) measures. Current guidance for best practices suggests that full psychometric review may be necessary to verify the underlying conceptual framework of the reformatted instruments[4]. This study of the SF-36v2<sup>®</sup> Health Survey (SF-36v2) examined the impact of changes in formatting necessary to create a simplified single item presentation by comparing data collected in that framework to data collected using a traditional display, that included grid display items, for score equivalence, uniformity of measurement properties and adherence to the conceptual framework of the SF-36v2.

Figure 1: Single Item Display of SF-36v2 Item VT04

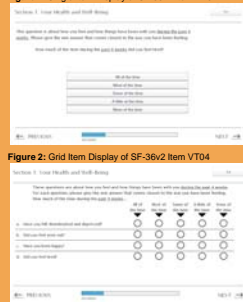
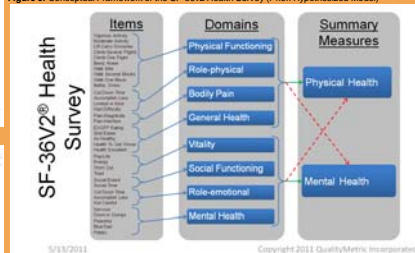


Figure 3: Conceptual Framework of the SF-36v2 Health Survey (Prior Hypothesized Model)



## METHODS

**Design:** QualityMetric, Incorporated conducted a survey of the U.S. general population. Data were collected for several self report health status measures in two waves during the spring and fall of 2009. Respondents were randomly assigned to one of four survey forms, only two of which apply to the current study. Forms varied either in terms of visual display, survey instrument or recall period. The two forms included in this study presented the SF-36v2 either in an item (single item or "SI") display or as a traditional grid display ("GD") (see Figures 1 & 2). The GD format was designed to simulate the traditional paper form of the survey. Response data were compared between SI and GD to investigate differences in survey properties between the two display formats.

**Instruments:** The U.S. English version of the SF-36v2 standard recall form was administered electronically. The SF-36v2 was developed in 1998 as a psychometrically superior replacement for the original version 1 of the survey. The SF-36v2 measures functioning and well-being across eight domains of health, and presents a two dimensional summary of results as Physical and Mental Component scores (PCS and MCS). Item stubs, health domains and their relations to summary measures are represented in Figure 3 above. In addition to the SF-36v2, a set of standard demographic items and a chronic conditions checklist were also presented to each respondent.

## ABSTRACT

**Objective:** Over the past two decades use of the traditional paper-and-pencil survey has waned as options for electronic data collection have been shown to be rigorous and more cost-effective. Although research supports equivalence of paper and electronic administration modes, evidence examining the impact of changes in item format required to accommodate small format electronic devices is lacking. This study examined the impact of a single item (SI) presentation versus grid display (GD) for score equivalence, measurement properties and adherence to the conceptual framework of the SF-36v2.

**Methods:** The SF-36v2 standard recall form was electronically presented as part of a US national norming study. Survey results from SI (N=2037) and GD (N=2003) administrations were then scored. ANCOVA models compared SI and GD scale scores. A Multi-trait Analysis Program (MAP-R) and principal components analysis (PCA) were used to examine the measurement properties and test the conceptual framework of SI and GD data.

**Results:** Mean score comparisons revealed small differences between SI and GD on seven scales (all p<.01). However, mean differences (.43 to 1.42) failed to reach the minimally important difference of 3 points indicating relative equivalence. MAP-R analyses showed that, for both item formats, SF-36v2 items had excellent convergent validity with their hypothesized scale (r>0.4) and each item correlated higher with its hypothesized scale than with others (divergent validity). PCA results showed that the hypothesized two-dimensional structure of physical and mental health was evident in both formats as the pattern of correlation between scales and principal components was consistent with a priori hypotheses and the two components explained the majority of variance in the eight scales (>75%).

**Conclusion:** SI presentation, which separates items from the contextual cues of their traditional grid format, results in scores and measurement properties consistent with GD, and maintains the underlying conceptual framework of the SF-36v2.

## METHODS (Continued)

**Sample:** Four thousand forty (4040) English speaking participants age 18 and over who answered one of two forms of the SF-36v2 as part of QualityMetric's 2009 PRO Norming Study were included for analysis in this project. Participants were selected using a methodology of pre and post weighting from Knowledge Networks' KnowledgePanel<sup>®</sup> which covers 97% of U.S. households. (See <http://www.knowledgenetworks.com/knpanel/KNPanel-Design-Summary.html> for KnowledgePanel information)

**Analyses:**  
• Descriptive statistics were calculated to identify sample characteristics. Chi-Square tests of difference were used to compare frequencies between SI and GD forms. SF-36v2 domains and summary scores were calculated using 2009 normative scoring.  
• ANCOVA models compared SI and GD domain scale scores and included demographic variables as covariates.  
• A scaling program (MAP-R) was used to examine item properties and their relationship to their domains.  
• Principal components analysis (PCA) was used to verify the conceptual framework of SI and GD data.  
• Differential item Functioning analyses were conducted to examine whether the perception of items was different between forms. (NOTES: Analyses updated since abstract to reflect the newest scoring and final cohort; Sample size may vary due to missing responses.)

## RESULTS

**Sample Characteristics:** Sample characteristics are reported in Table 1. Respondents completing the SI and GD forms were equivalent for all comparisons of sample characteristics indicating that the sampling paradigm was effective for creating equivalent cohorts.

**Mean Comparisons:** ANCOVA models were used to test the effect of form on SF-36v2 domain scores, holding constant the effects of sample characteristics (See Table 2). Mean score comparisons revealed small differences between SI and GD on three scales. However, the mean differences (0.11 to 1.35) failed to reach the minimally important difference of 3 points (0.3 of one SD) which has previously been established as appropriate for SF-36v2 summary and domain scale scores. This finding supports relative equivalence between the SI and GD presentation in terms of mean score.

Table 1: Sample Demographic Characteristics

	SI (2037)	GD (2003)
<b>Maies % (N)</b>	50.0 (1019)	48.7 (976)
<b>Age % (N) <math>\chi^2_{(5)} = 0.74, ns.</math></b>		
18-29	15.6 (318)	15.6 (310)
30-44	21.4 (436)	22.2 (444)
45-59	27.0 (549)	27.5 (550)
60+	36.0 (734)	34.9 (699)
<b>Education % (N) <math>\chi^2_{(5)} = 0.17, ns.</math></b>		
< High School	8.3 (168)	8.7 (174)
High School	30.1 (613)	30.4 (608)
Some College	31.5 (641)	30.3 (607)
>= Bachelors	30.2 (615)	30.7 (614)
<b>Ethnicity % (N) <math>\chi^2_{(5)} = 0.34, ns.</math></b>		
White NonHisp	77.3 (1574)	76.6 (1535)
Black NonHisp	8.1 (166)	9.2 (184)
Other NonHisp	4.0 (81)	4.0 (81)
Hispanic	9.6 (196)	10.1 (203)

Table 2: Domain Score Mean Differences and F-Test Coefficients

Domain	Mean Diff	F
Physical Func.	0.18	0.83
Role Physical	0.11	0.27
Bodily Pain	0.36	1.39
General Health	0.19	0.41
Vitality	0.24	0.73
Social Func.	0.72	5.59*
Role Emotional	0.84	7.30*
Mental Health	1.35	18.12**

Notes:  
\* Advantage GD  
\*\* p<0.05; \*p<0.01; \*\*p<0.0001

## RESULTS (Continued)

**Scaling Analysis:** MAP-R analyses revealed that, for both forms, SF-36v2 items had excellent convergent validity with their hypothesized scale (r >= .40 with scale). In addition, the SI form showed 100% divergent validity (each item correlated more highly with their hypothesized scale than with other scales) and GD form 97% (See Table 3).

**Principle Components Analysis:** PCA revealed a two component solution best fit the data from both forms. PCA results showed that the hypothesized two-dimensional structure of physical and mental health (See Figure 3) was evident in both the SI and GD formats, as the pattern of correlation between scales and principal components was consistent with a priori hypothesis [5,6] and the two components explained the majority of variance in the eight scales, each explaining more than 75% (see Table 4).

**Differential Item Functioning:** Logistic methods were utilized to assess DIF between forms. No DIF was detected for any item within any scale.

Table 3: Item Characteristics (Convergent and Divergent Validity) by Form

Domain	SI FORM			GD FORM		
	Min r	Max r	% Divergent Items	Min r	Max r	% Divergent Items
Physical Func.	0.58	0.85	100.0	0.58	0.85	90.0
Role Physical	0.80	0.93	100.0	0.86	0.91	100.0
Bodily Pain	0.80	0.80	100.0	0.77	0.77	100.0
General Health	0.48	0.79	100.0	0.44	0.73	100.0
Vitality	0.69	0.76	100.0	0.65	0.72	100.0
Social Func.	0.75	0.75	100.0	0.72	0.72	100.0
Role Emotional	0.82	0.92	100.0	0.83	0.88	100.0
Mental Health	0.65	0.79	100.0	0.61	0.76	100.0

Table 4: PCA Loading for Physical and Mental Components (Maximas rotation)

	SI Form (N=2031)			GD Form (N=1985)		
	Loadings	h2	PCs	Loadings	h2	PCs
Physical Functioning	0.90	0.20	0.85	0.91	0.17	0.85
Role Physical	0.88	0.31	0.87	0.88	0.31	0.87
Bodily Pain	0.78	0.33	0.73	0.73	0.34	0.65
General Health	0.59	0.52	0.62	0.54	0.55	0.59
Vitality	0.39	0.73	0.69	0.32	0.80	0.74
Social Functioning	0.48	0.73	0.77	0.55	0.64	0.72
Role Emotional	0.31	0.79	0.73	0.58	0.57	0.67
Mental Health	0.13	0.93	0.88	0.14	0.92	0.87

## CONCLUSIONS

This study examined the effects of changes in the display of SF-36v2 items by comparing data collected with our new single item display format with that collected with a more traditional format more closely replicating the paper form.

Based on this work we conclude:

- SF-36v2 items displayed using either our single item or traditional grid formats may be used without changing underlying item characteristics or the comparability of derived domain and summary scores.
- The underlying conceptual framework of the SF-36v2 is maintained in the two tested display formats.

**Limitations:** While significant efforts were made to select equivalent groups for participation in this study, some variation is attributable to differences in study respondents. In addition, the SI display was presented on a full-size computer screen and we have not yet tested this form of the SF-36v2 on a small screen devices.

**Implications:** This research suggests that the use of single-item administration forms of the SF-36v2 are appropriate for data collection as part of e-PRO protocols, and that the resulting data are equivalent to that collected by a grid based form providing evidence to satisfy requirements set forth by regulatory agencies.

## REFERENCES & NOTES

1. Food and drug Administration, Guidance for Industry, Patient-Reported Outcomes Measures: Use in Medical Product Development Support Labeling Claims (FDA, Rockville, MD, December 2009).
2. European Medicines Agency, Reflection Paper on the Regulatory Guidance for the use of Health Related Quality of Life (HRQL) Measures in the Evaluation of Medical Products (European Medicines Agency, London UK, July 2005).
3. Gwaltney CJ, Shields AL, & Shiffman S. Equivalence of Electronic and Pencil Administration of Patient-Reported Outcome Measures: A Meta-Analytic Review. Value Health 2008; 11(2): 322-333.
4. Coons SJ, Gwaltney CJ, Hays RD, et al. Recommendations on Evidence Needed to Support Measurement Equivalence Between Electronic And Paper-Based Patient-Reported Outcome (PRO) Measures: ISPOR ePRO Good Research Practices Task Force Report. Value Health 2009; 12(4): 419-426.
5. Ware JE, Kosinski M & Keller SD. SF-36 Physical and Mental Health Summary Scales: A User's Manual. Boston, MA: Health Assessment Lab, 1994.
6. Ware JE, Jr., Kosinski M, Bayliss MS, McHorney CA, Rogers WH, Raczek A. Comparison of methods for the scoring and statistical analysis of SF-36 health profile and summary measures: Summary of results from the Medical Outcomes Study. Medical Care 1995;33(Supplement 4):264-79.

© SF-36v2<sup>®</sup> Health Survey © 1992, 1996, 2000 Medical Outcomes Trust and QualityMetric Incorporated. All rights reserved. SF-36<sup>®</sup> is a registered trademark of Medical Outcomes Trust.  
\*\*\* This study was funded in its entirety by QualityMetric Incorporated.  
Presented at the 16<sup>th</sup> Annual International Society for Pharmacoeconomics and Outcomes Research, May 21<sup>st</sup>-25<sup>th</sup>, 2011, Baltimore, MD