Measurement Equivalence of Interactive Voice Response and Paper Versions of the EQ-5D in a Cancer Patient Sample

J. Jason Lundy, PhD*, Stephen Joel Coons, PhD
Patient-Reported Outcome Consortium, Critical Path Institute, Tucson, AZ, USA

ABSTRACT

Objective: To assess the measurement equivalence of an interactive voice response (IVR) version of the EQ-5D with the original paper version. Methods: Subjects were randomly assigned to: 1) paper then IVR, or 2) IVR then paper and asked to complete the questionnaire two days apart. The analyses tested mean differences (repeated measures analysis of variance) and reliability (intraclass correlation coefficient [ICC]). Equivalence of the means was established if the 95% confidence interval (CI) of the mean difference was within the minimally important difference interval, –0.035 to 0.035 for the EQ-5D index and –3 to 3 for the visual analog scale (EQ VAS). ICC adequacy was tested by comparing the ICC 95% lower CI with a critical value of 0.70. Results: The analyses included 113 subjects for the index and 109 subjects for the EQ VAS. For the index, the adjusted means of the paper and IVR versions were 0.789 ± 0.016 and 0.798 ± 0.017, respectively. The 95% CI of the mean difference was –0.024 to 0.006, within the equivalence interval. The ICC was 0.894 (95% lower CI 0.857), significantly greater than 0.70. For the EQ VAS, the adjusted means were 71.94 ± 1.87 for paper and 74.63 ± 1.79 for IVR. The 95% CI of the mean difference was –4.347 to –1.049, partially within the equivalence interval. The ICC was 0.887 (95% lower CI 0.840), significantly greater than 0.70. Conclusions: The results provide evidence that the EQ-5D scores on the IVR version were sufficiently equivalent to those obtained on the paper version.

Keywords: electronic data capture, EQ-5D, equivalence, interactive voice response, measurement IVR.

Copyright © 2011, International Society for Pharmacoeconomics and Outcomes Research (ISPOR). Published by Elsevier Inc.
timal based on the International Society for Pharmacoeconomics and Outcomes Research ePRO Task Force’s recommendations regarding the evidence needed to support measurement equivalence [13]. The Task Force report recommends the use of a crossover design with statistical tests of the mean differences and intraclass correlation coefficients (ICC). The purpose of this research was to assess the measurement equivalence of the original paper-based version and an IVR version of the EQ-5D, a widely used PRO measure [14], in a manner consistent with the recommendations of the ePRO Task Force.

Methods

Subjects
A nonprobability sample of subjects was obtained at the Arizona Cancer Center outpatient clinics. To qualify for inclusion, potential subjects were at least 18 years of age and currently in treatment (for cure or palliation) for cancer. Treatment included chemotherapy, radiation, a combination of both, or other medical treatments. Recruiting from among cancer patients who were undergoing treatment was intended to allow greater generalizability to patient populations participating in cancer clinical trials. Further, the subjects must have had access to and the ability to use a touch-tone telephone, as well as an understanding of written and spoken English.

Over a 6-month period (December 2007 through May 2008), project staff recruited potential subjects from the waiting areas of the cancer center’s medical and radiation oncology outpatient clinics. Recruitment materials emphasized the importance of this type of research in enabling clinicians and researchers to more effectively obtain patients’ perspectives regarding the impact of cancer and its treatment on their lives. Interested individuals had the opportunity to enroll based upon face-to-face contact with a recruiter at the clinics. Also, individuals who learned about the study from flyers were able to enroll by calling a dedicated telephone line. The study was conducted under the auspices of the University of Arizona’s Human Subjects Protection Program. All subjects who agreed to complete the study questionnaires received a $20 gift card.

Study design
A randomized crossover design was utilized in this study. Respondents were randomly assigned to complete either a paper questionnaire or the IVR-based questionnaire for the first administration and then the other mode for the second administration. Testing and order effects are threats to the validity of this design, but the within-patient design (i.e., subjects as their own control) provides greater statistical power and decreases sample size requirements. Therefore, the time between administrations should be adequate to minimize testing effects, or carryover, from the first administration, but not so long that the underlying domains being measured might actually change. Because the sample consisted of cancer patients receiving active treatment, the administration interval was chosen to be 48 to 72 hours. Depending on the stability of the constructs being measured, a retest interval between 2 and 14 days is generally used [15]. Our specific administration interval was selected to minimize the potential for the patient’s underlying health status to change, which would threaten the validity of this study design.

After random assignment, a suitably prepared study packet was mailed to each participant. The packet contained a cover letter, a study information sheet, the study disclaimer form, and two sealed envelopes labeled “1” and “2.” Each of these two envelopes was labeled with the date on which they were to be opened for self-administration of the questionnaire. The two administrations were scheduled, in consultation with the subject, two days apart during a relatively stable phase of their course of treatment. The subjects were contacted on the day of the first scheduled administration to remind or confirm completion of the first questionnaire and encourage them to complete the second. Each envelope contained information regarding completion of the designated mode of administration. The envelope instructing participants to complete IVR administration included written information regarding accessing the IVR system using their home telephone. The envelope instructing participants to complete the self-administered paper version of the questionnaire contained a self-addressed postage-paid return envelope. The cover letter asked the subject to complete the paper questionnaire and mail it back upon completion.

If there was a delay in receipt (via mail) or completion (via IVR) of the questionnaire project staff called to inquire about any difficulties the subject may be encountering and/or to confirm that the subject was fully aware of the study protocol timelines and encourage completion of the tasks. Each subject self-reported the date of completion of the paper questionnaire, whereas the IVR system captured the exact time and date of completion of the IVR questionnaire. The recorded dates were compared against the scheduled completion dates for confirmation that the questionnaire was completed within the 72-hour administration window.

Study measures
During the recruitment and enrollment process, information was obtained from subjects regarding year of birth, sex, cancer diagnosis (i.e., cancer site/type), and type of treatment (i.e., chemotherapy, radiation, or a combination of both). The questionnaire, which included the EORTC QLQ-C30 along with the EQ-5D, took about 10 to 15 minutes to complete each time.

EQ-5D
The EQ-5D self-report questionnaire consists of two parts: a descriptive system and a visual analog scale [14]. The descriptive system comprises the following five dimensions: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. For each dimension, respondents are asked to describe their own health with respect to three levels, reflecting “no problems,” “some problems,” and “extreme problems.” These responses are used to classify the respondent into one of 243 unique EQ-5D health states. A scoring function assigns a value to self-reported health states from a set of preference weights that have been empirically derived [16]. The resulting EQ-5D index score is on a scale where 1.0 represents perfect health and 0.0 represents death.

In addition to the multidimensional descriptive system, the EQ-5D has a visual analog scale (EQ VAS) to measure the respondent’s overall self-assessed health status. The original EQ-VAS is a thermometer-like 20-cm vertical line with endpoints labeled “worst imaginable health state” and "best imaginable health state" anchored at 0 and 100, respectively. The EQ-5D descriptive system was adapted to the IVR using the exact wording for the items and responses. For the EQ-VAS, the IVR system asked the respondent to “picture in your mind” a scale with 100 at the top (i.e., “best health state you can imagine”) and 0 at the bottom (i.e., “worst health state you can imagine”) and enter a number between 0 and 100 that reflects his or her health status “today.”

Data analysis
Descriptive statistics of cancer type, age, and sex were calculated to characterize the sample; no subgroup analyses were performed. All statistical analyses of the EQ-5D data were performed using SPSS version 16.0 (Chicago, IL) and evaluated using a one-sided alpha (α) level of 0.05.
Mean differences. Testing of the mean differences was based on analysis of variance (i.e., split-plot analysis of variance) with factors for mode, period of administration (first or second), and subject; the $p$ values from the significance tests will be reported. The split-plot analysis of variance also accounts for the interaction effect (period $\times$ mode effect, often called carryover). We included the period effect since it accounts for learning or any other period effects as described in Hills and Armitage [17] and because we had no a priori evidence that it could be ignored. The adjusted mean differences between modes were estimated together with the associated 95% confidence interval (CI) for the difference. Equivalence on this measure was considered to have been established if the 95% confidence interval excludes the minimally important difference (MID) used in the sample size calculations, namely 0.07 for the EQ-5D index and six points for the EQ VAS. Thus, the equivalence intervals were $-0.035$ to $+0.035$ for the EQ-5D index and $-3.0$ to $+3.0$ for the EQ VAS [18,19].

Reliability. To analyze the reliability or reproducibility of the measurement between modes, the analyses were based upon the ICC [20]. The ICC is calculated based on the analysis of variance model that includes factors for mode and subject; the analysis is the same regardless of whether mode is treated as a fixed or random effect. A one-sided 95% CI for the lower bound was computed using the formula provided in McGraw and Wong [21]. Measurement equivalence was considered to have been established if the lower bound of the 95% CI exceeded 0.70 [22].

Sample size considerations
The sample size for this study was based on two different calculations providing the sample size for testing mean differences as well as determining the magnitude of the ICC. The sample size calculations used for the test of mean differences comes from Lachin [23], whereas the sample size for the ICC analyses is based on the estimation of the magnitude of the reliability coefficient from Streiner and Norman [15]. The resulting sample sizes based on the calculations in Lachin [23], for a 95% CI, were 50 subjects for the EQ-5D index and 70 subjects for the EQ VAS. The sample size calculation to test the ICC from Streiner and Norman [15] was 108 subjects. It is important to note that the sample size calculations provide the number of completed pairs necessary for the desired statistical power.

The more conservative sample size calculation, which is provided from the computation of the sample requirements for our tests of the ICC, was used as the target sample size. Opting for the higher sample size provided sensitivity around the Lachin [23] calculations and also provided a large enough sample size to allow for dropouts and incomplete data. Hence, we targeted 110 subjects for recruitment into this study.

Results

Sample characteristics
A total of 184 subjects agreed to participate. Of those, 139 subjects completed both administrations for a response rate of 75.5% (see Figure 1). This sample of respondents was 67.6% female and had a mean age of 61.5 years. The ages ranged from 19 to 86 years. Furthermore, 37.4% ($n=52$) of this sample reported a diagnosis of breast cancer, 17.3% ($n=24$) reported a diagnosis of lung cancer, 11.5% ($n=16$) reported a diagnosis of colorectal cancer and 7.9% ($n=11$) reported a diagnosis of prostate cancer. The remaining cancer types ($n=36$) included eight cases of melanoma, five cases each of ovarian cancer and lymphoma, and 18 diagnoses of other cancer types. The predominant treatment for this sample was chemotherapy ($n=100$, 71.9%) whereas 12.2% were receiving radiation therapy ($n=19$) and eight subjects had undergone surgery (5.8%). All other subjects ($n=12$) were receiving other treatment strategies (e.g., hormones).

Data quality
The amount of missing and unusable data from the paper and IVR administrations was tabulated for all subjects who completed both questionnaire administrations. Of the 139 paper responses, only two missing responses, from the same subject, were noted on the descriptive system. Similarly, only two missing responses were observed, from different subjects, on the IVR version of the descriptive system. The IVR version allowed people to actively skip questions by pressing the 9 key on the telephone key pad. There were four subjects out of the 139 (2.9%) who did not complete the paper EQ VAS in contrast to no missing data on the IVR version of the EQ VAS. Further, there were 53 subjects (38.8%) who did not complete the paper EQ VAS according to the instructions. Twenty-four subjects (17.3%) drew a line from the bottom of the VAS scale (i.e., originating from “0”) to the point representing the valuation of their health. Thirteen subjects (9.4%) drew a circle around the corresponding VAS response and an additional 13 subjects drew a line across the VAS to select their value. Three other unusual responses were noted: two subjects drew an arrow pointing to their VAS response and one subject placed an “X” on the scale.

Although no missing responses were noted on the IVR version of the EQ VAS, this mode was not free of unusual phenomenon. There were eight subjects who entered a single-digit response on the IVR EQ VAS. Four of those single-digit responses corresponded exactly with the first digit of the two-digit VAS response given on paper. For example, a subject may report an EQ VAS score of 70 on the paper version but the IVR version score was recorded as a score of seven. Two subjects reported paper VAS scores of 70 and 40 but the corresponding IVR scores were eight and five, respectively. Of the remaining two subjects, one subject did not respond to the paper VAS and the other subject had very disparate scores of a 40 on the paper and a five on the IVR version.

Equivalence testing
The analyses for the index score and the EQ VAS were based on a per protocol analysis. Hence, subjects were in the analyses if their score difference exceeded two standard deviations (95% of the score distribution), namely 40 points on the EQ VAS
Discussion

The results of the comparison of the paper version of the EQ-5D with the IVR version provide substantial evidence supporting the measurement equivalence of these two modes of administration. However, these results do have limitations, including a study sample that may lack generalizability to the general population, the cancer patient population, and patients with other conditions or diseases. Because the internal validity of measurement equivalence studies is of greater importance, we felt this was an acceptable tradeoff. For measurement equivalence studies, the preservation of internal validity to detect evidence of the introduction of systematic bias is of foremost importance. Although the possibility of these findings being sample-dependent remains, the choice of sample is unlikely to have a large impact on determining whether bias was introduced when migrating from paper to IVR. Therefore, whereas the magnitude of the scores is not generalizable to other populations, the lack of systematic differences between the paper and IVR version of the EQ-5D may transcend the sample from which it resulted. In addition, because only a US-English language version of the EQ-5D was used in this study, measurement equivalence of other language or cross-cultural translations of the EQ-5D remains unassessed.

Further, we encountered a small amount of unusable data from the IVR version of the EQ VAS. We observed the phenomenon of a subject entering a single digit (i.e., “8”) upon one administration and having a value of 70 on the other administration. We believe it is unlikely that subjects intended such a wide disparity when evaluating their current health status, particularly over the 2- to 3-day interval used in this study. Although the disparate values were only observed in eight subjects (5.7%), this situation resulted in unreliable data provided by the electronic platform. To overcome this limitation, we recommend that when items with response sets exceeding single digits (e.g., EQ VAS) are adapted to IVR systems, that the system prompts the subject to confirm his or her choice. This solution could be easily incorporated and should eliminate this problem.

Finally, although not a part of our original analytic plan, we conducted analyses that included all completed score pairs within the 72-hour administration window for the EQ-5D index and EQ VAS. Due to the inclusion of additional sources of variance, the results demonstrate lower levels of score agreement; however, they do not change the conclusions of this study. The results reinforce the need to carefully specify the per protocol criteria, a priori, in an attempt to limit sources of variability that may result from study design considerations when testing measurement equivalence.

By using a crossover design and tests of mean differences and ICCs, our study and analytical strategy conformed to the recommendations of the International Society for Pharmacoeconomics and Outcomes Research ePRO Task Force regarding the evidence needed to support measurement equivalence [13]. The mean difference CI for the EQ-5D index score reflected equivalence of the means from the two modes, as the CI was wholly contained inside the equivalence interval. The mean difference CI for the EQ VAS was only partially contained in the equivalence interval, providing an inconclusive result of either equivalence or nonequivalence of the mean EQ VAS scores. However, it is worth noting that others

<table>
<thead>
<tr>
<th>Table 1 – Mean differences and intraclass correlation coefficients (ICCs) for the EQ-5D scores.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Adjusted means (SE)</strong></td>
</tr>
<tr>
<td>Paper EQ-5D index</td>
</tr>
<tr>
<td>IVR EQ-5D index</td>
</tr>
<tr>
<td>Paper EQ VAS</td>
</tr>
<tr>
<td>IVR EQ VAS</td>
</tr>
</tbody>
</table>

CI, confidence interval; IVR, interactive voice response; SE, standard error; VAS, visual analog scale. * Adjusted means are equivalent to least squares means.

and 0.28 on the EQ-5D index score. The analysis included 109 subjects for the EQ VAS and 113 subjects for the index. Of the 30 EQ VAS exclusions, 21 subjects were beyond the 72-hour completion window, four subjects had missing EQ VAS scores, and five subjects had missing index scores. Of the 26 subjects excluded from the EQ-5D index analysis, 22 subjects were beyond the 72-hour completion window, two subjects had missing index scores, and two subjects had scores considered outliers.

The means of the paper and IVR administrations of the EQ-5D index were 0.790 ± 0.172 and 0.800 ± 0.180, respectively. The tests for an order effect and for an order by mode interaction based on the split-plot analysis of variance were not statistically significant. The adjusted means (i.e., least squares means) were 0.789 ± 0.016 for the paper version and 0.798 ± 0.017 for the IVR version. The adjusted mean difference was −0.009 and the 95% CI of the mean difference was −0.024 to 0.006, which was within the equivalence interval. The ICC was 0.894 (95% lower CI 0.857), significantly greater than 0.70 (Table 1). Furthermore the percent of exact agreement and kappa coefficients for each of the five dimensions of the descriptive system are provided in Table 2.

The EQ VAS means were 71.9 ± 19.7 for paper and 74.6 ± 18.7 for IVR. There was no order effect present; however, there was a significant mode by order interaction in the analysis of means for the EQ VAS (P = 0.022). The adjusted means were 71.94 ± 1.87 for the paper and 74.63 ± 1.79 for the IVR. The adjusted mean difference was −2.69 ± 0.83 and associated 95% CI of the mean difference was −4.347 to −1.049, partially contained within the equivalence interval of −3 to +3. The ICC was 0.887 (95% lower CI 0.830) also significantly greater than 0.70 (Table 1).

To assess the robustness of our “per protocol” analyses, we included subjects who were removed because their score difference exceeded two standard deviations between administrations. We did not include those who were outside the administration window of 72 hours. Due to the nature of our cancer patient sample, a change in the underlying health condition of the subjects was a key threat to the internal validity of the study. Hence, the administration window was specified a priori to mitigate this effect and ensure a stable sample for equivalence testing. In an analysis of all completed score pairs within the 72-hour window on the EQ-5D index (n = 115), the adjusted mean difference was 0.014 (−0.002 to 0.030) and the ICC was 0.873 (95% lower CI 0.822). A similar analysis of EQ VAS responders (n = 114) yielded an adjusted mean difference of 0.53 (−1.99 to 3.05) and an ICC of 0.803 (95% lower CI 0.727).

<table>
<thead>
<tr>
<th>Table 2 – Percent agreement and kappa coefficients for the five EQ-5D dimensions.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Percent exact agreement</strong></td>
</tr>
<tr>
<td>EQ-5D dimension</td>
</tr>
<tr>
<td>Mobility</td>
</tr>
<tr>
<td>Self-care</td>
</tr>
<tr>
<td>Usual activities</td>
</tr>
<tr>
<td>Pain/discomfort</td>
</tr>
<tr>
<td>Anxiety/depression</td>
</tr>
</tbody>
</table>

SE, standard error.
have operationalized measurement equivalence intervals differently in the literature [24,25]. For instance, in a study assessing the measurement equivalence of English and Chinese versions of the EQ-5D, Luo et al. [24] used smaller MIDs than we did for the EQ-5D index (0.05 vs. 0.07) and the EQ VAS (5.0 vs. 6.0); however, they operationalized their equivalence intervals as −0.05 to +0.05 for the EQ-5D index and −5.0 to +5.0 for the EQ VAS. Hence their equivalence intervals were double their MIDs, whereas our equivalence intervals were equal in magnitude to our chosen MIDs. In addition, both the index and EQ VAS scores in our study had ICCs or kappa coefficients that were on par or higher than those observed in the literature for the test–retest comparison of the paper version.

As stated in the ePRO Task Force recommendations, electronic modes of administration should not be held to a higher standard than the original paper-based version [13]. Although limited, some data are available regarding within-mode (i.e., paper to paper) test–retest results for the EQ-5D. Macran [26] reports reliability coefficients for the EQ-5D descriptive system (i.e., dimensions) and the EQ VAS for two different test-retest paper administrations with a 3-month interval. The first study used data collected in the Measurement and Valuation of Health study and reported kappa coefficients that ranged from 0.49 to 0.75 on the descriptive system and an ICC of 0.84 on the EQ VAS. The second analysis from a rheumatoid arthritis study reported kappas on the five dimensions that ranged from 0.40 to 0.65 and an ICC of 0.78 for the EQ VAS. The kappas from our study were mostly higher and from a smaller range, albeit from a much shorter time interval. The ICCs for the EQ VAS were higher than those reported by Macran [26]. Moreover, there were substantial percent-ages of absolute agreement, above 79% on all dimensions, between the paper and IVR versions of the EQ-5D.

Conclusions

This study compared the scores derived from paper and IVR versions of the EQ-5D in a sample of cancer patients. The evidence presented here, when taken in totality, supports the measurement equivalence of the IVR version of the EQ-5D with the original paper version. Investigators should have confidence that the scores obtained from the IVR version of the EQ-5D are comparable to those obtained on the paper version.

Acknowledgments

The authors thank Robin Whitacre, Vern Hartz, and the recruitment team at the Arizona Cancer Center’s Behavioral Measurements Shared Service for their assistance. The authors also thank the Perceptive Informatics Team, particularly Stephanie Clive, Steven Flaherty, and Hannah O’Gorman for IVR delivery support and Damian McIntegart, James Pierce, and Keith Wenzel for study design and implementation support.

Source of financial support: This project was funded by a grant to the University of Arizona from Perceptive Informatics (ClinPhone PLC, at the time of this study). The authors have no financial interest in Perceptive Informatics or ClinPhone PLC. The data used in this research were collected as part of a study funded by Perceptive Informatics (formerly ClinPhone PLC). Additional support was provided by the Arizona Cancer Center Support Grant CA023074 from the National Cancer Institute. The authors were employed by the University of Arizona at the time this study was conducted. The views expressed in this manuscript are those of the authors and do not necessarily reflect the position or policy of the University of Arizona or Critical Path Institute.

REFERENCES