Volume 11 • Number 2 • 2008
VALUE IN HEALTH

Equivalence of Electronic and Paper-and-Pencil Administration of Patient-Reported Outcome Measures: A Meta-Analytic Review

Chad J. Gwaltney, PhD,[1,5] Alan L. Shields, PhD,[2,5] Saul Shiffman, PhD[3,4,5]

[1]Brown University, Providence, RI, USA; [2]East Tennessee State University, Johnson City, TN, USA; [3]University of Pittsburgh, Pittsburgh, PA, USA; [4]invivodata, Inc., Pittsburgh, PA, USA; [5]PRO Consulting, Pittsburgh, PA, USA

ABSTRACT

Objectives: Patient-reported outcomes (PROs; self-report assessments) are increasingly important in evaluating medical care and treatment efficacy. Electronic administration of PROs via computer is becoming widespread. This article reviews the literature addressing whether computer-administered tests are equivalent to their paper-and-pencil forms.

Methods: Meta-analysis was used to synthesize 65 studies that directly assessed the equivalence of computer versus paper versions of PROs used in clinical trials. A total of 46 unique studies, evaluating 278 scales, provided sufficient detail to allow quantitative analysis.

Results: Among 233 direct comparisons, the average mean difference between modes averaged 0.2% of the scale range (e.g., 0.02 points on a 10-point scale), and 93% were within ±5% of the scale range. Among 207 correlation coefficients between paper and computer instruments (typically intraclass correlation coefficients), the average weighted correlation was 0.90; 94% of correlations were at least 0.75. Because the cross-mode correlation (paper vs. computer) is also a test–retest correlation, with potential variation because of retest, we compared it to the within-mode (paper vs. paper) test–retest correlation. In four comparisons that evaluated both, the average cross-mode paper-to-computer correlation was almost identical to the within-mode correlation for readministration of a paper measure (0.88 vs. 0.91).

Conclusions: Extensive evidence indicates that paper- and computer-administered PROs are equivalent.

Keywords: computer, electronic, equivalence, meta-analysis, paper and pencil, patient-reported outcomes.

Patient-reported outcome (PRO) measures—i.e., self-reported measures of health status—are increasingly being used in medical and drug development studies [1–4]. PRO data are valuable for several reasons: 1) for many outcomes (e.g., pain, depression), patient reports are the best available method for obtaining information on unobservable events; 2) even when an event is observable (e.g., voiding, dietary intake), the patient is often in the best position to assess and report these outcomes; 3) PRO measures may be more reliable and valid than measures completed by a clinician via an interview with the patient; and 4) in the case of health-related quality of life measures, PRO data can uniquely provide information on a patient's perception of both a disorder and the treatment for the disorder. In sum, PRO measures supply valuable information on health status and treatment effects that could not be collected in any other way.

The use of computers to collect PRO data is becoming commonplace. Computerized assessments potentially offer a number of advantages over paper and pencil assessments [5]: 1) missing data within an assessment can be reduced by requiring completion of an item before the patient can move on to the subsequent question; 2) computerized assessments can handle complex skip patterns, which often confound patients and result in incomplete or invalid data [1]; 3) computerized assessments eliminate out of range and ambiguous data by allowing the patient to only select one of the on-screen response options; 4) computerized assessments reduce the effort and error involved in entering paper PRO data; 5) in diary studies, electronic diaries can implement sophisticated designs to ensure valid representation of the patient's experience [6]; and 6) electronic data capture can time-tag records to document timely compliance and can increase compliance. Compliance with computerized diaries is often 90% or better, whereas studies have documented only 11% to 20% compliance with paper diaries [7,8]. Thus, there are several reasons why clinicians and researchers may prefer computerized administration to paper and pencil PRO measures.

Despite its promise, the shift to electronic patient-reported outcomes (ePROs) requires establishing the equivalence of PRO measures administered on a

Address correspondence to: Chad J. Gwaltney, Brown University, Box G-S121-5, Providence, RI 02912, USA. E-mail: chad_gwaltney@brown.edu

10.1111/j.1524-4733.2007.00231.x

  1098-3015/08/322 322–333

computer and the original paper and pencil versions [9,10]. In other words, evidence may be needed to demonstrate that scores derived from a computerized measure do not differ from scores derived from the paper and pencil version. Given that the computer and paper versions of PROs present the same text content and response options, one might expect them to be equivalent. Nevertheless, there are two primary reasons why computerized measures might not be equivalent: 1) differences in how the items and responses are presented to the respondent; and 2) potential difficulties that some individuals may have in interacting with computers. The first category encompasses a number of changes that are required to present a PRO measure on a computer. These changes can range from very minor changes, such as asking the patient to tap a response on a computer screen instead of circling a response on a page, to substantial changes, such as splitting items and responses onto multiple screens, because of space constraints [11]. A common change is that items are presented on a computer one at a time, although multiple items are generally presented on the same page in a paper and pencil assessment. This could alter responding if the participant refers to previous questions when answering the current item (e.g., referring back to one's responses about symptom intensity when considering overall health state). Although computerized assessments usually allow participants to move back through an assessment to view or change previous items, it does make the process more difficult, which could influence responding. Assessments can be implemented on different platforms with varying screen sizes, ranging from small-screen personal digital assistants (PDAs) to large screen desktop computers. Because of the smaller screen size, more changes to the presentation of the assessment items may be required with PDAs, which could alter responding. Although migrating an assessment to a computer could adversely affect the instrument, there is also some evidence that computerized assessments can result in more valid data, especially when "sensitive" topics, such as drug use or risky sexual behaviors, are targeted [12,13].

The second concern touches on characteristics of the patient that may impede responding to an assessment completed on a computer. For example, individuals with high levels of "computer anxiety" might report more negative mood when completing a mood assessment on a computer [10,14]. More broadly, patients with little computer experience might have more difficulty completing computerized measures, resulting in a nonequivalent measure.

In this article, we assess the equivalence of ePRO assessments to their paper ancestors. The American Psychological Association (APA) [9] defines equivalence as a demonstration that: 1) the rank orders of scores of individuals tested in alternative modes closely approximate each other; and 2) the means, dispersions, and shapes of the score distributions are approximately the same. Many empirical studies have addressed the equivalence of computerized PRO measures and paper and pencil PRO measures. Consistent with APA guidelines, most studies of PRO equivalence assess the correlation and/or mean differences between computerized and paper and pencil measures. In most cases, this is accomplished using a crossover design, where a patient completes one version of the PRO measure and later completes the other version. Ideally, the order of computer and paper administration is randomized to control for possible order effects. Studies using intraclass correlations, which account for both covariation (as do other correlations) and equivalence of means and variance [15], provide a particularly strong assessment of equivalence.

In this article, we use meta-analysis to summarize results from studies quantifying the relationship between computerized and paper measures. Possible moderators were also considered. Migrating an assessment to a computer may differentially impact the responding of older patients. Therefore, studies including an older patient sample might exhibit lower correlations between paper and computerized measures and greater mean differences. We also examined the possibility that studies enrolling patients with less computer experience would exhibit larger mean differences and lower correlations. Further, we addressed whether the platform (PDA vs. PC) on which the computerized assessment was administered influences equivalence.

Examining the equivalence of paper and computerized measures is essentially an examination of test–retest or alternate-forms reliability. This sets a high bar for demonstrating equivalence. Correlations between the two modes of administration should not only be high and significant, but should also meet requirements for demonstrating reliability. A test–retest correlation of 0.75 or higher is considered "excellent" [16,17] and was used as the standard of comparison here. It is also important to place the correlation in the context of test–retest correlations for two administrations of the paper measure. Variations in scores between paper and ePRO can occur either because of random variation or because of changes in the construct between assessments, which would also affect repeat administrations of a paper measure. Accordingly, we also compared the paper-to-ePRO correlations with test–retest correlations for paper. If the correlations are similar, this would be very strong evidence for the equivalence of the paper and computer measures.

Methods

Identification and Selection of Studies

We conducted a computerized literature review using the PsycInfo (1887–2006) and PubMed (1966–2006)

databases. In the literature review, we located articles by crossing the keywords *quality of life, symptom, depression, anxiety, mood*, and *pain* with the keywords *computer, electronic, paper*, and *web*. A manual search was also performed by scanning reference lists of reviewed articles, to identify other articles that were not selected in our computerized search. The literature review returned 65 studies that assessed equivalence.

The eligibility criteria for inclusion of a selected study in the meta-analysis were: 1) published in English; 2) published in a peer-reviewed journal or conference proceedings; and 3) presented correlations or measures of agreement (intraclass correlation coefficients [ICCs], Pearson product-moment correlations, Spearman rho, weighted kappa) and/or mean differences between paper and computer measures. (A study that used unweighted kappas was excluded, because unweighted kappa is an overly conservative method for determining agreement between scales with more than two response options.) Of the 65 studies assessing equivalence, 46 met these inclusion criteria (as coded independently by CG and AS, with discrepancies resolved by discussion). Studies often included analyses of multiple measures (Table 1); there were 233 codable mean difference scores from 38 studies and 207 codable correlations from 32 studies.

Migrating assessments from paper to electronic platforms invariably requires changing the assessment presentation to some degree. Although details on the changes made to the electronic versions were rarely described in detail, most studies appeared to be "faithful migrations," in which the electronic items very closely resembled the paper-and-pencil items. Systematic changes appeared to include changes in font size, orientation of the response scale (e.g., a horizontal response array was presented vertically on the electronic version), and presenting a single item on screen at a time (vs. having multiple items on the same page in a pencil and paper measure). There was no indication that more substantial changes, such as changes to the item wording or response scales, had been made in any of the studies.

Coding of Analyses and Moderators

The type of correlation coefficient used varied across studies. The ICC was the most commonly used correlation (Table 1). Like all conventional correlations, the ICC ranges from –1 to 1. Nevertheless, unlike conventional correlations, it not only evaluates the strength of association between two variables, it also assesses the degree of equivalence between score distributions [18]. In the context of this article, high positive ICCs indicate not only that paper and computer measures covary, but also that the mean and variability of the scores are similar. The Pearson product-moment correlation and Spearman rho were also used in some studies. One study used weighted kappa, a measure of agreement that is similar to the ICC [19], to assess equivalence.

Although a few studies also presented correlations and/or mean differences for individual items within scales, we focused on equivalence of scales, because the individual items are rarely used as stand-alone measures in clinical trials. Nevertheless, where a single item is the entire measure (e.g., a visual analog pain scale), the single item correlations and mean differences were included in analysis. In addition to recording correlations and mean differences, we also coded (when available) the average age of the sample included in the study, degree of computer experience, and the platform on which the computerized assessment was administered. One study did not present the average age of the sample, but did note that the minimum age was 65 years. Therefore, 65 was coded as the average age. A single study on children (average age 13 years) was excluded, because all of the other studies assessed adults. A minority of studies assessed computer experience and the scale used to measure experience varied across studies. The most commonly reported scale (n = 9) was frequency of use. Therefore, we coded the percent of the sample that never or rarely used computers, e.g., less than 1x/month. Computer platform was dichotomized as PDA versus larger screen device. We did not assess studies of interactive voice response systems (IVRS), both because few studies were available and, more importantly, because migration of a written scale to auditory scale involves much more substantive changes (i.e., aural presentation of items and responses, serial presentation of content, and so on), which might substantially diminish equivalence.

Computation and Analysis of Summary Indices

Two methods were used to summarize the equivalence data. A correlation is a measure of effect size (ES) and meta-analysis was used to synthesize these results [20]. (The weighted kappas used in one study were included in analysis with the correlation coefficients. When we use the term "correlations," we refer to all the variations of correlation coefficients and the weighted kappa.) Data were aggregated using a weighted linear combination, giving greater weight to studies with larger sample sizes. The Comprehensive Meta-Analysis, v2 software was used to aggregate the correlations. In calculating the summary correlation across studies (the ES from this point forward), multiple correlations within one study (e.g., multiple scales assessed in the same study) were averaged, so that a single study did not disproportionately contribute to the meta-analysis. Analyses of the disaggregated correlations produced results that were very similar to the analyses using the aggregated data and are not presented here. Using the disaggregated data did not

Table I Characteristics of studies included in meta-analysis

								Equivalence indices				
									Correlation			
Study	Design	Random	eMode	N	Pop	Measure	No. of scales tested	Mean differences*	ICC	PPM	SR	WK
[28]	C	N	PC/Laptop	43	Rheumatology	HAQ; Psych Distress; Pain VAS; Fatigue VAS; Overall VAS	5	−2.4	0.92			
[29]	C	Y	PC/Laptop	30	Rheumatology	WOMAC	3	−3.1	0.91			
[30]	C	Y	PC/Laptop	50	Rheumatology	BAS; Quebec scale	4	−2.0	0.92			
[31]	C	Y	PC/Laptop	53	Rheumatology	WOMAC	3	−0.6	0.94			
[32]	C	Y	PC/Laptop	26	Cardiology	SAQ	5	−1.2	0.78			
				29		SF-36	8					
[33]	C	Y	PC/Laptop	138	Psychiatry	SF-36	9		0.88			
[34]	C	Y	PC/Laptop	51	Asthma (adults)	AQLQ	5	−0.4	0.93			
				52	Asthma (pediatric)	PAQLQ	4					
[35]†	C	Y	PDA	35 (p)	GI	IBS-QOL; EQ-5D	11	0.2	0.91			
				23 (p)		WPAI	3					
				35 (p)		WPAI	1					
				37 (e)		IBS-QOL; EQ-5D	15					
				27 (e)		WPAI	4					
[36]	C	Y	PDA	47	Alcohol	Alcohol effects	3	1.3				
[37]	C	N	PDA	68	Asthma	SF-36; AQLQ	13	−0.4	0.96			
[38]	C	Y	PC/Laptop	189	Pain	SF-MPQ; PDI	6	0.3			0.76	
[39]	C	Y	PDA	40	GI	PGWB	7	−1.3				
[40]	P	Y	PC/Laptop	52	Psychiatry	STAI	2	0.2				
[41]	C	Y	Tablet	38	Cancer	CCM	7			0.93		
[42]	C	Y	PDA	24	Cancer	Pain (Avg and max)	2	1.4				
[43]	P	Y	PC/Laptop	97	College students	BDI; STAI	3	3.5				
[44]	C	N	PC/Laptop	29	Post-natal	EPDS	1	−0.7		0.98		
[45]	C	N	PC/Laptop	32	Rheumatology	VAS (Pain, Fatigue, Arthritis)	3	1.0	0.90			
		Y		40		RQLQ	1					
[46]	C	Y	PDA	40	Asthma	Asthma VAS	1	5.8		0.86		
[47]	C	Y	PC/Laptop	43	Healthy controls	CESD	1	−0.8				
[48]	C	N	PDA	20	Pain	Pain VAS	1			0.84		
[49]	C	Y	PDA	24	Healthy volunteers	Pain VAS	1			0.97		
[50]	C	Y	PC/Laptop	134	GI	QOLRAD	6	0.1	0.91			
[51]	C	N	PC/Laptop	54	Elderly primary care	CESD-R; GDS; ADL; IADL	4	−0.3		0.73		
[52]	C	N	PDA	30	Rheumatology	Pain VAS; Fatigue VAS; Global VAS; RADAI; MHAQ; SF-36	20	0.7		0.91		
[53]	C	Y	PC/Laptop	75	Allergy	RQLQ; WPAI	8	1.7			0.90	
[54]	C	Y	PC/Laptop	66	Mood	STAI; BDI	2	−0.2				
[55]	P	N	PC/Laptop	3247	Alcohol	Multiple indicators of alcohol consumption frequency	7	−0.3				
[56]	C	N	PC/Laptop	130	Rheumatology	ACRPA Pain/ Overall VAS	2	−2.0	0.83			
[57]	C	N	PC/Laptop	113	National screening day	CESD	1				0.96	
[58]	P	Y	PDA	60	Pain	Pain (Avg and max)	2	−0.7				
[59]	C	Y	PC/Laptop	76	Diabetics	WBQ; DTSQ	6	1.6			0.78	
[60]	C	Y	PC/Laptop	115	Pain	SF-36	8	0.4				
[61]	C and P	N	PDA	87	Rheumatology	SF-36; WOMAC	11	1.1	0.86			
[62]	C	Y	PC/Laptop	50	Rheumatology	NASS	2	1.4	0.94			
[63]	P	N	PC/Laptop	64	Psychiatric Clinic	SCL90	10	1.5				
[64]	P	Y	PC/Laptop	196	Psychiatric Clinic	SCL90	10	1.7				
[65]	C	N	PC/Laptop	10	Psychiatry	Beck Anxiety Inventory	1	−1.1		0.93		
[66]	C	Y	PDA	12	Appetite	EARS	6	−0.7				
[67]	C	N	PDA	20	Appetite	EARS	8			0.91		
[68]	C	Y	PC/Laptop	50	Cancer	EORTC	9	−1.0		0.90		

Table 1 continued

Study	Design	Random	eMode	N	Pop	Measure	No. of scales tested	Equivalence indices: Mean differences*	Correlation: ICC	PPM	SR	WK
[69]	C	Y	PC/Laptop	50	Rheumatology	WOMAC	3	–0.6	0.89			
[70]	C	N	PC/Laptop	53	Rheumatology	WOMAC Pain	4	1.3				
[71]	C	N	PC/Laptop	88	Urology	IPSS	1			0.90		
[72]	C	Y	PC/Laptop	149	Cancer	EORTC; HADS	11	0.2				0.67
[73]	C	N	PC/Laptop	51	Rheumatology	SF-36	8			0.91		

*Mean differences are expressed as a percent of the scale range. Note that the numbers in this column are average scores over all mean differences reported in each study.
†Bushnell et al. 2006 include two sets of correlations and mean differences: one set when the paper version was administered first and then the electronic version, and a second set when the administration order was reversed. The (p) and (e) in the sample size column indicates which administration came first for those analyses. The sample size for the WPAI differed from the rest of the questionnaires, requiring separate rows. Further, in the set of results where the paper version was administered first, the sample used for analysis of one of the WPAI scales differed from the sample used for the other three scales.
Design: C, crossover; P, parallel; Random (For crossover designs—Was order of administration randomly selected? For parallel designs—Were participants randomly assigned to groups?): Y = yes; N = no; ICC, Intraclass Correlation; PPM, Pearson Product Moment Correlation; SR, Spearman's rho; VAS, Visual Analog Scale; WK, Weighted Kappa; WPAI, Work Productivity and Activity Impairment Questionnaire.

alter the ES, but did make it more reliable and statistically significant (because of the increased number of comparisons). A random effects model was used to calculate the pooled ES estimate, although a mixed effects model was used to examine the influence of moderators on the ES (where the moderator is treated as a fixed effect). Analyses using a fixed-effect model produced similar results and are not reported here. Publication bias or the "file-drawer effect" was examined using the Orwin Fail-Safe N [21]. The Fail-Safe N estimates how many missing studies would need to be added to the analysis, to bring the overall ES below a specified level. It also allows the researcher to specify the mean ES of the missing studies. As the number of required missing studies increases, confidence in the ES estimate also increases. We estimated the number of missing studies that would be required to bring the overall ES below 0.75. Heterogeneity among the study correlations was assessed using the Qw statistic. Potential outliers were identified using residual plots and dropped from analysis, if appropriate. Studies using ICC were compared to studies using other correlations. Because the ICC takes both covariance and score distribution into account, similarity between the ICC and other correlations provides additional evidence for the equality of the scores.

The difference between means is also an ES, but cannot be analyzed by meta-analysis without being standardized in some way, traditionally by a measure of the pooled standard deviation of scores. However, the pooled standard deviation was infrequently included in the published articles. Because it is not possible to interpret the meaningfulness or clinical significance of absolute differences between modes of administration, we expressed the mean difference as a percentage of the response scale (e.g., the range of a 0–100 scale was coded as 101); e.g., if the mean difference was 5 scale points and the scale range was 100, the standardized mean difference was coded as 5%.

We examined whether platform type influenced the magnitude of the correlations by calculating an ES for each type of platform and then examining differences between these subgroupings using the Qb statistic [20]. For age and computer experience, we ran a meta-regression analysis to determine whether these continuous variables were linearly related to ES [22].

To account for the fact that analyses are nested within studies, generalized estimating equations (GEE [23]); were used to examine the relationship between the moderator variables and the mean difference scores. Because the size of the mean differences was of primary interest, not the direction, the absolute mean difference scores were used as the primary dependent variable in analysis. For age, a positive association indicates that mean differences get larger as the sample gets older. For computer experience, a positive association indicates that as the percent of the sample with little computer experience increases so do the mean differences. Electronic platform is a categorical variable (PDA vs. larger screen, including desktop and laptop). GEE was used to examine whether the mean differences varied as a function of the type of platform used.

Results

Study Characteristics

Characteristics of all studies included in the meta-analysis are listed in Table 1. The number of analyses for a single study ranged from 1 to 20. The average age of participants in the studies was 48.0 ± 13.9. Among studies reporting computer experience, the average percentage of the sample that never or rarely used computers was 39.4% ± 23.4. Thirty percent of the studies used a PDA for computerized assessments.

Overall Relationship between Paper and Computerized Assessments

Mean differences. Among the 233 mean differences evaluated, the average mean difference was 0.2% of the scale range. In other words, on a 100-point scale, the mean of the scores from a computerized measure

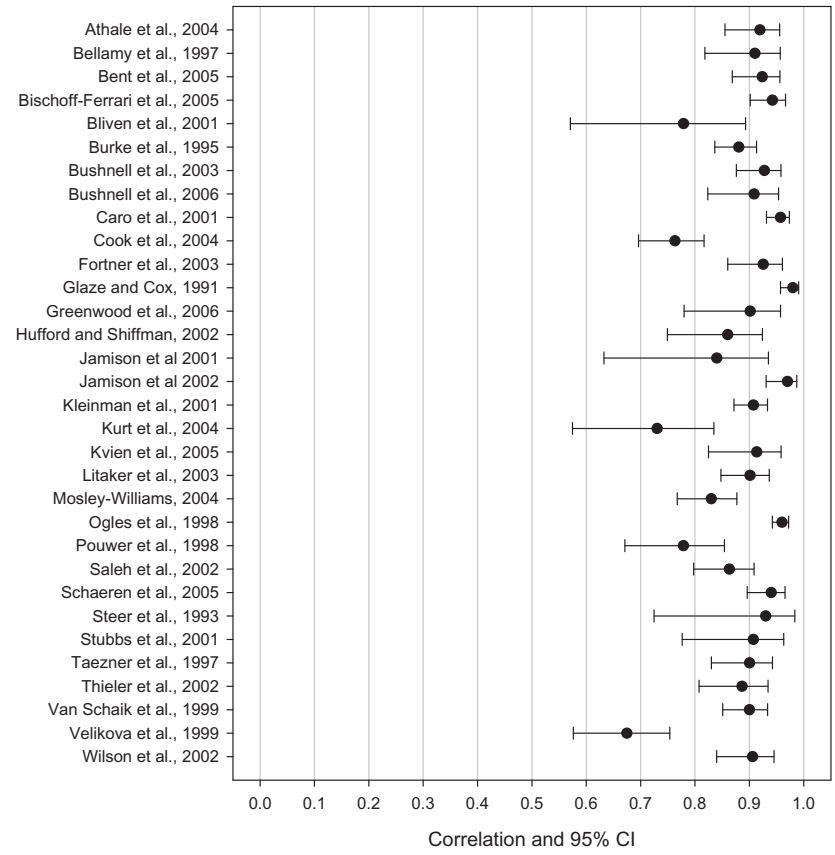

Figure 1 Correlations from each study.

was 0.2 points higher than the mean of the scores from a paper and pencil measure. The average mean difference did not significantly differ from 0, $t(232) = 0.93$, ns. The mean differences ranged from –7.8% to 7.6% of the scale score. The mean difference was within ±5% of the scale score in 93% of the studies.

To ensure that the small value for the mean difference between modes was not due to negative and positive differences canceling out, we also evaluated the absolute mean differences. The average absolute mean difference was 2.0%; on a 100-point scale, the average difference between electronic and paper measures is 2 points.

Correlations. Correlations from each study are shown in Figure 1. Of the 32 studies contributing correlations, 30 (94%) had average correlations that were greater than 0.75 (89% of the disaggregated correlations were also ≥0.75). The weighted summary correlation was 0.90, 95% CI 0.87–0.92. Studies using ICC or weighted kappa produced ES that were exactly the same as studies using Pearson or Spearman correlations (0.90), Qb = 0.07, ns.

To gauge the effects of possible publication bias ("file-drawer effect") and the possibility that relevant studies were not identified in our literature review, we calculated the number of missing studies that would be required bring the estimated average correlation below 0.75, using the Orwin fail-safe N [21]. The average correlation for the missing studies was specified at 0.68, the lowest correlation observed among the published studies. Using this criterion, 95 additional studies would be required to reduce the ES below 0.75. Even if the missing studies had, on average, a moderate correlation of 0.50, it would require 32 additional studies to reduce the ES below 0.75.

There was statistically significant variability among the ES estimates, Qw (31) = 215.8, $P < 0.001$, suggesting that the effect varies across studies [20]. Therefore, we identified studies with extreme ES by calculating the standardized residual score for each study (standardized difference between each study ES and the weighted mean ES). Eight studies had absolute residual scores of three or greater (four had correlations above the mean and four below the mean). Dropping these studies from analysis eliminated the heterogeneity among the scores, Qw (23) = 32.3, ns, and the estimated correlation was unchanged, 0.90, 95% CI 0.89–0.91, $P < 0.001$ for both fixed and random effects models.

The "outlier" studies were reviewed to identify factors (PRO measure used, sample size and characteristics, platform, study design) that might explain their extreme correlations, but no commonalities or patterns

were observed. Because 1) excluding the outlier studies did not influence the summary ES; 2) there is no methodological characteristic that explains the deviant correlations; and 3) the absolute differences among the ES are not particularly large [20], these studies were retained in the analyses.

Test–Retest Reliability of Pencil and Paper Measures

There were four studies (encompassing 44 scales) that examined both paper-computer concordance and paper-paper test–retest reliability. In these studies, the correlation between the paper and computer scores (average 0.88, 95% CI 0.85–0.91) was very similar to the test–retest reliability of the paper measure (average 0.91, 95% CI 0.86–0.94). These correlations did not differ significantly, Qb (1) = 0.83, ns. This demonstrates that even the modest observed variation between paper and electronic forms is not due to changing modalities, but to random variation across multiple administrations.

Analysis of Moderator Effects

Mean differences. We used GEE to examine the relationship between platform and the mean differences. The raw mean differences for both types of platform are small and not significantly different from 0, PDA: M = 0.7% of scale range, SD 2.0, 95% CI –0.7–2.2%; larger-screen platforms: M = –0.1%, SD 1.5, 95% CI –0.7–0.5%. Initial analysis of the effect of platform type suggested a difference in absolute means, contrast coefficient = –0.81, SE = 0.35, $P < 0.05$. Although initially significant, this relationship was extremely small (0.7% of the scale range: absolute mean differences from PDA studies = 2.4% of scale range, from larger screen studies = 1.7%), and dissipated when a single outlier study was excluded (contrast coefficient = –0.51, SE = 0.35, ns.). (The outlying study was unique in that it used a relatively long interassessment interval [24 hours]. It also included measures—the Work Productivity and Activity Impairment Questionnaire (WPAI) and EuroQoL (EQ-5D)—that may have required substantial modifications in migrating to PDA: some WPAI items have extensive introductory text, which could be difficult to fit on a PDA screen, and the EQ-5D has a 20-cm Visual Analog Scale (VAS), which would need to be dramatically shortened on a PDA screen, which may explain why it differed from the others.)

The association between computer experience and the absolute mean differences was not significant, linear coefficient = 0.002, SE = 0.006, ns. Age was also unrelated to the size of the mean differences, linear coefficient = 0.001, SE = 0.01, ns.

Correlations. Comparing studies that used a PDA (n = 8) versus studies that used a larger screen device (n = 23), there were no differences in correlation with paper Qb (1) = 0.50, ns. The average correlation for PDA studies was 0.91 (95% CI 0.87–0.94), and the average correlation for larger screen studies was 0.90 (95% CI 0.86–0.92). Accounting for platform did not resolve the heterogeneity of the ES; there was significant heterogeneity among the ES within each platform subgroup.

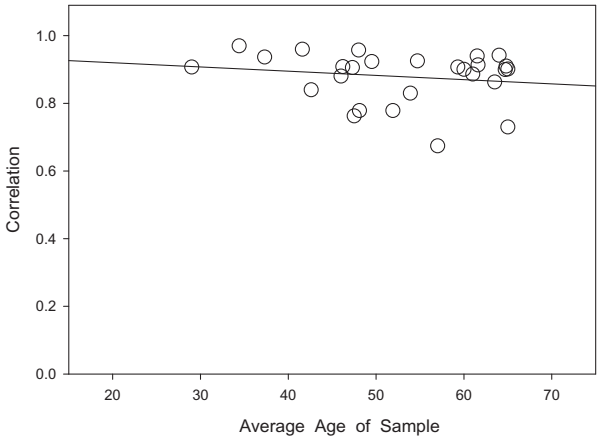

Figure 2 Scatterplot and regression line for association between age and paper-electronic correlation.

Age was significantly associated with the paper-electronic correlations (transformed to Fisher Z′), such that the correlations decreased as age increased (Fig. 2), slope estimate = –0.008, SE = 0.003, $P < 0.01$. Nevertheless, this trend was quite small: with each 1-year increase in the average sample age, the Fisher Z′ decreases by 0.008 points. In other words, with each decade of age, the Fisher Z′ decreases by about 0.08 (equivalent to a change in r of 0.02). The correlations were generally in the acceptable range (>0.75) among even the oldest samples (Fig. 2).

The percent of participants who had never or rarely used computers was unrelated to the correlations, slope estimate = –0.004, SE = 0.002, ns.

Conclusions

The results summarized here show that computer and paper measures produce equivalent scores. Mean differences were very small and neither statistically nor clinically significant. Correlations were very high, and were similar to correlations between repeated administrations of the same paper-and-pencil measure.

Administering PRO measures on computer has the potential to improve patient compliance and reduce the data management burden on investigators. Nevertheless, it has been suggested that investigators need to evaluate equivalence when a PRO measure is moved from paper to electronic administration. For example, the FDA Draft Guidance on PRO endpoints suggests that migrating a measure from paper to computer

requires validation testing to ensure that the computerized measure is equivalent to the paper measure [1]. We reviewed the substantial literature on the subject to assess the equivalence of paper and electronic administration. The data from almost 300 comparisons yield an unambiguous conclusion: paper and pencil and computerized measures produce equivalent scores.

According to APA guidelines [9], one method for demonstrating equivalence is to examine differences between the average scores derived from the different administration modes. Mean differences between the two modes of administration were small—the average difference was only 0.2% of the scale range. There were very few instances where the difference exceeded 5% of the scale range. In a particular application, the investigator must evaluate differences associated with method of assessment relative to clinically meaningful "minimally important differences" [24]. Although we could not evaluate the observed differences in relation to the minimally important difference, which differs across measures and populations, the observed mean differences—less than one half point on a 100-point scale—appear to be so small as to be of no practical significance in any context. Moreover, the meta-analysis showed that the mean difference was not significantly different from zero, indicating that even this small-observed difference was likely due to random variation. Thus, the mean differences were very small and suggest equivalence.

In addition to mean differences, it is also important to examine the correlation between scores from each administration mode, to determine whether individuals retain their relative "rank" in the score distribution when completing the computerized measure [9]. In these data, the average weighted correlation was 0.90, suggesting that relative position in the distribution is retained when the assessment is completed on a computer. Further, studies using ICC or weighted kappa, which take into account both covariance and score means and variability, yielded equivalence estimates that were almost identical to studies using more traditional correlation coefficients. This provides compelling evidence that there is little change in patient responses when migrating to an electronic platform.

There was substantial heterogeneity in the individual correlations, which was resolved by dropping several outliers, without affecting the overall ES. Nevertheless, we were unable to identify any methodological factors within the outlier studies that explained their extreme ES. The analysis of moderators did not resolve this heterogeneity. Therefore, it is unclear why these studies produced extreme correlations. Some factor that we could not identify may cause the variation, or it could be due to random variability in a distribution. In fact, four of the studies were above the mean and four below the mean, as would be expected in a normal distribution. Even with this heterogeneity, only two of the studies produced a correlation that was less than 0.75. There is no reason to believe that the unexplained heterogeneity should temper the overall conclusions drawn from the meta-analysis.

When an assessment is completed on different occasions, there are many reasons why scores may change—among other things, participants may change their minds about how to respond to an item, their actual condition may change even when the interval between assessments is short, or simple random error may alter responses. This is why even repeated administrations of the same test, in the same modality, vary; yielding test–retest correlations lower than 1.0. Because the equivalence tests of paper and electronic assessments also involved two administrations, some of the variability in scores between paper and electronic tests is due to this test–retest variation, rather than to mode of administration. To assess how much of the observed variation was due to the changes in mode of administration, we compared the observed paper-to-electronic correlations to the test–retest correlations from two administrations of the paper assessment. The two measures of retest variation were very similar: in other words, administering a test on a computer is just like readministering the paper test a second time. This suggests that there is little or no variation in scores attributable to mode of administration, and provides compelling evidence that the computerized measures are equivalent to the paper and pencil versions.

The meta-analysis showed a high overall level of agreement between paper and computerized measures. We also examined whether equivalence varied by age, computer experience, and computer platform. Even though items may need to be altered to fit on a PDA screen (e.g., reducing the size of a VAS [11]), there is little evidence that using a PDA decreases concordance with the paper version of the assessment. Mean differences were slightly larger in studies using PDAs, but this effect was small and largely due to one outlier study. Additionally, paper-computer correlations were not moderated by platform type—correlations were above 0.90 for both types of platform. Most importantly, the mean differences were not significantly different from zero for either type of platform, suggesting that both PDAs and larger screen devices produce scores that are equivalent to scores from paper forms. There was no variation by subjects' computer experience. Although increasing age was associated with lower paper-electronic correlations, this association was small and unlikely to be clinically relevant. Even when the average age of the sample is approximately 65 years old, the predicted paper-electronic correlation is 0.86. It is also possible that test–retest reliability is adversely affected by increasing age, regardless of mode of administration [25], which would explain the observed slight decline in paper-computer concordance.

The uniform results seen in this review have implications for the use of computerized measures in clinical trials: as long as substantial changes are not made to the item text or response scales, equivalence studies should not be necessary to demonstrate anew the equivalence or validity of a computerized measure. The studies we reviewed appeared to use "faithful migrations," where the exact text of the paper instrument was ported to a computer screen, without making substantive changes in content. However, a limitation of this literature is that little information is provided about alterations that were made to the items to present them via a computer. For example, studies using PDAs would almost certainly have made some minor revision of items (e.g., placing general instructions on an introductory screen, followed by individual items), but these details are generally not reported. Our finding of equivalence cannot be directly generalized to cases where substantial changes are made to item content or where layout changes substantially affect users' ability to respond to the item, such as when questions are separated from response options or when scrolling is required to view an entire item [11]. When substantive changes like these have been made to a computerized measure, equivalence studies such as those reviewed here may be necessary.

Although equivalence testing should not be needed in most cases when migrating an assessment from paper to computer, it may be fruitful to evaluate the changes in formatting, layout, etc., through cognitive interviewing techniques to ensure that the patients are interpreting the items as intended. Cognitive interviewing is a qualitative method for assessing respondents' interpretation of the assessment, using a small sample of patients studied in the laboratory [26]. Through these small-scale studies, it is possible for investigators to determine whether the alterations made in migrating to computer influence the way in which the assessment is understood by patients.

Our conclusions cannot be generalized to all forms of electronic administration of PROs. We specifically addressed the case of written assessments moved from paper to computer administration. None of the studies reviewed here used an IVRS as the mode of electronic assessment. Studies examining IVRS have assessed the equivalence of IVRS measures with clinician-administered assessments (see review in [27]), not assessments completed directly by the patient. Those studies suggest the equivalence of IVRS and interview measures. However, the equivalence of IVRS and clinician-administered assessments does not indicate that IVRS measures are also equivalent to paper and pencil measures completed by the patient, nor can the conclusions of our review of written measures be generalized to IVRS. IVRS measures are fundamentally different from written measures, in that: 1) they are presented aurally, not visually; 2) the information is presented serially and the patient is required to retain the question text and response categories in working memory as the item is presented; 3) subjects cannot review the item or response array at a glance; and 4) whereas responses on a computer screen are typically presented in a meaningful order that helps subjects place themselves in the response set (e.g., from low to high severity), responding on a telephone keypad may disrupt this ordered physical representation of responses. Because clinician-administered assessments share these characteristics, it is not surprising that they are equivalent to IVRS measures. Nevertheless, because of the substantial differences between IVRS and patient-completed PRO measures, further testing is required before concluding that they are equivalent.

We have demonstrated that written assessments administered on paper and by computer are equivalent. This suggests that scores obtained via the two modalities are directly comparable. This finding should be doubly reassuring to investigators using electronic PROs in the context of randomized trials, where the focus is on comparison across groups that both use electronic assessment (thus making equivalence to paper instruments less of an issue).

The use of computerized measures to collect PRO data is likely to grow, because electronic assessment offers many advantages over paper and pencil measures. This growth need not be impeded by concerns about the equivalence of electronic PRO measures to their paper-and-pencil ancestors.

Source of financial support: No external funding was received to support the completion of this manuscript.

References

1 Food and Drug Administration. Guidance for industry: Patient Reported Outcome measures: use in medical product development to support labeling claims. Fed Regis 2006;71:1–32.
2 Wiklund I. Assessment of patient-reported outcomes in clinical trials. The example of health-related quality of life. Fundam Clin Pharmacol 2004;18:351–63.
3 Willke RJ, Burke LB, Erickson P. Measuring treatment impact: a review of patient-reported outcomes and other efficacy endpoints in approved product labels. Appl Clin Trials 2004;25:535–52.
4 Wilson IB, Cleary PD. Linking clinical-variables with health-related quality-of-life—a conceptual-model of patient outcomes. JAMA 1995;273:59–65.
5 Richard DCS, Lauterbach D. Computers in the training and practice of behavioral assessment. In: Haynes SN, Heiby EM, eds. Comprehensive Handbook of Psychological Assessment, Vol. 3: Behavioral Assessment. Hoboken, NJ: John Wiley & Sons, 2004.
6 Stone AA, Shiffman S. Ecological momentary assessment (EMA) in behavioral medicine. Ann Behav Med 1994;16:199–202.

7 Hufford MR, Shields AL. Electronic subject diaries: an examination of applications and what works in the field. Appl Clin Trials 2002;11:46–56.
8 Stone AA, Shiffman S, Schwartz JE, et al. Patient non-compliance with paper diaries. BMJ 2002;324: 1193–4.
9 American Psychological Association. Guidelines for Computer-Based Tests and Interpretations. Washington, DC: American Psychological Association, 1986.
10 Schulenberg SE, Yutrzenka BA. The equivalence of computerized and paper-and-pencil psychological instruments: implications for measures of negative affect. Behav Res Methods Instrum Comput 1999; 31:315–21.
11 Palmblad M, Tiplady B. Electronic diaries and questionnaires: designing user interfaces that are easy for all patients to use. Qual Life Res 2004;13:1199–207.
12 Locke SE, Kowaloff HB, Hoff RG, et al. Computer-based interview for screening blood-donors for risk of HIV transmission. JAMA 1992;268:1301–5.
13 Kobak KA, Taylor LV, Dottl SL, et al. A computer-administered telephone interview to identify mental disorders. JAMA 1997;278:905–10.
14 Tseng HM, Tiplady B, Macleod HA, Wright P. Computer anxiety: a comparison of pen-based personal digital assistants, conventional computer and paper assessment of mood and performance. Br J Psychol 1998;89:599–610.
15 McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychol Methods 1996;1:30–46.
16 Cicchetti DV, Sparrow SS. Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior. Am J Ment Def 1981;86:127–37.
17 Fleiss JL. Statistical Methods for Rates and Proportions (2nd ed.). New York: Wiley, 1981.
18 Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979; 86:420–8.
19 Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educ Psychol Meas 1973;33: 613–19.
20 Hedges LV, Olkin I. Statistical Methods for Meta-Analysis. Orlando, FL: Academic Press, 1985.
21 Orwin RG. A fail-safe N for effect size in meta-analysis. J Educ Stat 1983;8:157–9.
22 Rosenthal R. Meta-analysis: a review. Psychosom Med 1991;53:247–71.
23 Zeger SL, Liang K, Albert PS. Models for longitudinal data: a generalized estimating equation approach. Biometrics 1988;44:1049–60.
24 Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. Control Clin Trials 1989; 10:407–15.
25 Grafton KV, Foster NE, Wright CC. Test-retest reliability of the Short-Form McGill Pain Questionnaire: assessment of intraclass correlation coefficients and limits of agreement in patients with osteoarthritis. Clin J Pain 2005;1:73–82.
26 Willis GB. Cognitive Interviewing: A Tool for Improving Questionnaire Design. Thousand Oaks, CA: Sage Publications, 2005.
27 Byrom B, Mundt JC. The value of computer-administered self-report data in central nervous system clinical trials. Curr Opin Drug Discov Devel 2005;8:374–83.
28 Athale N, Sturley A, Skoczen S, et al. A web-compatible instrument for measuring self-reported disease activity in arthritis. J Rheumatol 2004;31: 223–8.
29 Bellamy N, Campbell J, Stevens J, et al. Validation study of a computerized version of the Western Ontario and McMaster Universities VA3.0 Osteoarthritis Index. J Rheumatol 1997;24:2413–15.
30 Bent H, Ratzlaff CR, Goligher EC, et al. Computer-administered bath ankylosing spondylitis and Quebec Scale outcome questionnaires for low back pain: agreement with traditional paper format. J Rheumatol 2005;32:669–72.
31 Bischoff-Ferrari HA, Vondechend M, Bellamy N, Theiler R. Validation and patient acceptance of a computer touch screen version of the WOMAC 3.1 osteoarthritis index. Ann Rheum Dis 2005;64:80–4.
32 Bliven BD, Kaufman SE, Spertus JA. Electronic collection of health-related quality of life data: validity, time benefits, and patient preference. Qual Life Res 2001;10:15–22.
33 Burke JD, Burke KC, Baker JH, Hillis A. Test-retest reliability in psychiatric-patients of the SF-36 health survey. Int J Methods Psychiatr Res 1995;5:189–94.
34 Bushnell DM, Martin ML, Parasuraman B. Electronic versus paper questionnaires: a further comparison in persons with asthma. J Asthma 2003;40:751–62.
35 Bushnell DM, Reilly MC, Galani C, et al. Validation of electronic data capture of the Irritable Bowel Syndrome—Quality of Life Measure, the Work Productivity and Activity Impairment Questionnaire for Irritable Bowel Syndrome and the EuroQol. Value Health 2006;9:98–105.
36 Cameron E, Sinclair W, Tiplady B. Validity and sensitivity of a pen computer battery of performance tests. J Psychopharmacol 2001;15:105–10.
37 Caro JJ Sr, Caro I, Caro J, et al. Does electronic implementation of questionnaires used in asthma alter responses compared to paper implementation? Qual Life Res 2001;10:683–91.
38 Cook AJ, Roberts DA, Henderson MD, et al. Electronic pain questionnaires: a randomized, crossover comparison with paper questionnaires for chronic pain assessment. Pain 2004;110:310–17.
39 Drummond HE, Ghosh S, Ferguson A, et al. Electronic quality of life questionnaires: a comparison of pen-based electronic questionnaires with conventional paper in a gastrointestinal study. Qual Life Res 1995;4:21–6.
40 Ford BD, Vitelli R, Stuckless N. The effects of computer versus paper-and-pencil administration on measures of anger and revenge with an inmate population. Comput Human Behav 1996;12:159–66.
41 Fortner B, Okon T, Schwartzberg L, et al. The Cancer Care Monitor: psychometric content evaluation and

pilot testing of a computer administered system for symptom screening and quality of life in adult cancer patients. J Pain Symptom Manage 2003;26:1077–92.
42 Gaertner J, Elsner F, Pollmann-Dahmen K, et al. Electronic pain diary: a randomized crossover study. J Pain Symptom Manage 2004;28:259–67.
43 George CE, Lankford JS, Wilson SE. The effects of computerized versus paper-and-pencil administration on measures of negative affect. Comput Human Behav 1992;8:203–9.
44 Glaze R, Cox JL. Validation of a computerized version of the 10-item (self-rating) Edinburgh Postnatal Depression Scale. J Affect Dis 1991;22:73–7.
45 Greenwood MC, Hakim AJ, Carson E, Doyle DV. Touch-screen computer systems in the rheumatology clinic offer a reliable and user-friendly means of collecting quality-of-life and outcome data from patients with rheumatoid arthritis. Rheumatology (Oxford) 2006;45:66–71.
46 Hufford MR, Shiffman S. Correspondence between paper and electronic visual analog scales among adult asthmatics. Paper Presented at the Drug Information Association Statistics Conference. Hilton Head, South Carolina, 2002.
47 Izquierdo-Porrera AM, Manchanda R, Powell CC, et al. Factors influencing the use of computer technology in the collection of clinical data in a predominantly African-American population. J Am Geriatr Soc 2002;50:1411–15.
48 Jamison RN, Raymond SA, Levine JG, et al. Electronic diaries for monitoring chronic pain: 1-year validation study. Pain 2001;91:277–85.
49 Jamison RN, Gracely RH, Raymond SA, et al. Comparative study of electronic vs. paper VAS ratings: a randomized, crossover trial using healthy volunteers. Pain 2002;99:341–7.
50 Kleinman L, Leidy NK, Crawley J, et al. A comparative trial of paper-and-pencil versus computer administration of the Quality of Life in Reflux and Dyspepsia (QOLRAD) questionnaire. Med Care 2001;39:181–9.
51 Kurt R, Bogner HR, Straton JB, Tien AY, Gallo JJ. Computer-assisted assessment of depression and function in older primary care patients. Comput Methods Programs Biomed 2004;73:165–71.
52 Kvien TK, Mowinckel P, Heiberg T, et al. Performance of health status measures with a pen based personal digital assistant. Ann Rheum Dis 2005; 64:1480–4.
53 Litaker D. New technology in quality of life research: are all computer-assisted approaches created equal? Qual Life Res 2003;12:387–93.
54 Lukin ME, Dowd ET, Plake BS, Kraft RG. Comparing computerized versus traditional psychological assessment. Comput Human Behav 1985;1:49–58.
55 McCabe SE, Diez A, Boyd CJ, et al. Comparing web and mail responses in a mixed mode survey in college alcohol use research. Addict Behav 2006;31:1619–27.
56 Mosley-Williams A, Williams CA. Validation of a computer version of the American College of Rheumatology Patient Assessment questionnaire for the autonomous self-entry of self-report data in an urban rheumatology clinic. Arthritis Rheum 2004;50:332–3.
57 Ogles BM, France CR, Lunnen KM, et al. Computerized depression screening and awareness. Community Men Health J 1998;34:27–38.
58 Palermo TM, Valenzuela D, Stork PP. A randomized trial of electronic versus paper pain diaries in children: impact on compliance, accuracy, and acceptability. Pain 2004;107:213–19.
59 Pouwer F, Snoek FJ, van der Ploeg HM, et al. A comparison of the standard and the computerized versions of the Well-being Questionnaire (WBQ) and the Diabetes Treatment Satisfaction Questionnaire (DTSQ). Qual Life Res 1998;7:33–8.
60 Ryan JM, Corry JR, Attewell R, Smithson MJ. A comparison of an electronic version of the SF-36 General Health Questionnaire to the standard paper version. Qual Life Res 2002;11:19–26.
61 Saleh KJ, Radosevich DM, Kassim RA, et al. Robinson H. Comparison of commonly used orthopaedic outcome measures using palm-top computers and paper surveys. J Orthop Res 2002;20:1146–51.
62 Schaeren S, Bischoff-Ferrari HA, Knupp M, et al. A computer touch-screen version of the North American Spine Society outcome assessment instrument for the lumbar spine. J Bone Joint Surg Br 2005;87:201–4.
63 Schmitz N, Hartkamp N, Brinschwitz C, Michalek S. Computerized administration of the Symptom Checklist (SCL-90-R) and the Inventory of Interpersonal Problems (IIP-C) in psychosomatic outpatients. Psychiatry Res 1999;87:217–21.
64 Schmitz N, Hartkamp N, Brinschwitz C, et al. Comparison of the standard and the computerized versions of the Symptom Check List (SCL-90-R): a randomized trial. Acta Psychiatr Scand 2000;102:147–52.
65 Steer RA, Rissmiller DJ, Ranieri WF, Beck AT. Structure of the computer-assisted Beck Anxiety Inventory with psychiatric-inpatients. J Pers Assess 1993;60: 532–42.
66 Stratton RJ, Stubbs RJ, Hughes D, et al. Comparison of the traditional paper visual analogue scale questionnaire with an Apple Newton electronic appetite rating system (EARS) in free living subjects feeding ad libitum. Eur J Clin Nutr 1998;52:737–41.
67 Stubbs RJ, Hughes DA, Johnstone AM, et al. Description and evaluation of a Newton-based electronic appetite rating system for temporal tracking of appetite in human subjects. Physiol Behav 2001;72:615–19.
68 Taenzer PA, Speca M, Atkinson MJ, et al. Computerized quality-of-life screening in an oncology clinic. Cancer Pract 1997;5:168–75.
69 Theiler R, Spielberger J, Bischoff HA, et al. Clinical evaluation of the WOMAC 3.0 OA Index in numeric rating scale format using a computerized touch screen version. Osteoarthritis Cartilage 2002;10:479–81.
70 Theiler R, Bischoff-Ferrari HA, Good M, Bellamy N. Responsiveness of the electronic touch screen WOMAC 3.1 OA Index in a short term clinical trial with rofecoxib. Osteoarthritis Cartilage 2004;12: 912–16.
71 van Schaik P, Ahmed T, Suvakovic N, Hindmarsh JR. Effect of an educational multimedia prostate program

on the International Prostate Symptom Score. Eur Urol 1999;36:36–9.
72 Velikova G, Wright EP, Smith AB, et al. Automated collection of quality-of-life data: a comparison of paper and computer touch-screen questionnaires. J Clin Oncol 1999;17:998–1007.
73 Wilson AS, Kitas GD, Carruthers DM, et al. Computerized information-gathering in specialist rheumatology clinics: an initial evaluation of an electronic version of the Short Form 36. Rheumatology (Oxford) 2002;41:268–73.