ANALYSIS CONCEPT PROTOCOL

MODE EFFECTS OF PATIENT-REPORTED OUCTOMES DATA COLLECTION

1. STUDY TITLE

Mode effects of patient-reported outcomes data collection in childhood cancer survivors: A Report from Childhood Cancer Survivor Study

2. WORKING GROUP AND INVESTIGATORS

This proposed publication will be reviewed by the Psychology Working Group (Primary) and Biostatistics/Epidemiology Working Group (secondary). Proposed investigators will include:

I-Chan Huang	SJCRH	i-chan.huang@stjude.org	(901) 595-8365
Greg Armstrong	SJCRH	greg.armstrong@stjude.org	(901) 595-5892
Yutaka Yasui	UALBERTA	yutaka.yasui@ualberta.ca	
Wendy Leisenring	FHCRC	wleisenr@fhcrc.org	(206) 667-4374
Melissa Hudson	SJCRH	melissa.hudson@stjude.org	(901) 595-3445
Les Robison	SJCRH	les.robison@stjude.org	(901) 595-6078
Kevin Krull	SJCRH	kevin.krull@stjude.org	(901) 595-5891

Note:
SJCRH: St Jude Children's Research Hospital
UALBERTA: University of Alberta
FHCRC: Fred Hutchinson Cancer Research Center

3. BACKGROUND AND RATIONALE

For the last three decades, the 5-year survival rate of childhood cancer has improved substantially in the United States, from less than 50% in the 1970s to 80% today (1). It is estimated that the number of childhood cancer survivors will be approximately 420,000 by the end of 2013 (2), and 24% of them having survived greater than 30 years (3). Childhood cancer survivors are vulnerable to significant late effects (2). The Childhood Cancer Survivor Study (CCSS) has reported that 62% of adult survivors of childhood cancer had ≥1 chronic conditions (4), and by 50 years old 53% of survivors had developed a life threatening or fatal condition (5). Late effects can influence cancer survivors' health-related quality of life (HRQOL) (6,7), which is defined as perceived well-being and capability of performing daily functions as a result of treatment and/or disease (8,9).

CCSS cohort was conventionally investigated via mail survey as the primary mode and telephone interview as the secondary mode to collect self-reported health status, HRQOL, lifestyle and behavioral outcome data. In addition to the telephone and mail modes, the web-based mode is used in CCSS expansion cohort as well. Advances in electronic and mobile health technologies (eHealth and mHealth, respectively), such as web-based computer interfaces, handheld devices, and electronic data capture, have substantially improved collection of self-reported outcomes data. eHealth/mHealth possesses several advantages including the reduction of subject burden, decrease of incomplete data, avoidance of secondary data entry, accurate implementation of skip patterns, flexible administration in different locations (e.g., at home or in clinic), and automatic reminders for repeated data collection (10,11). Additionally, the web-based mode can include a variety of validation and skip sequence procedures. The migration from mail survey and telephone

interview to eHealth/mHealth-based survey is especially desirable for young adult survivors of childhood cancer as they more frequently engage with computer technologies and electronic devices.

One essential issue in migrating survey administration from the mailed questionnaires and telephone interviews to eHealth/mHealth mode is to assure minimal impacts on the measurement characteristics of the surveys. The International Society for Pharmacoeconomics and Outcomes Research (ISPOR) Electronic Patient-Reported Outcome (ePRO) Good Research Practices Task Force (2009) has called for the evidence of measurement equivalence when migrating from the paper-and-pencil mode to the electronic mode for collecting self-reported health outcome data (12). ISPOR PRO Mixed Modes Good Research Practices Task Force (2014) further recommends that collecting self-reported health outcomes data should not be varied within studies that seek to pool or compare the data without prior evidence of sufficient measurement invariance between the modes (13). This recommendation is based on the notion that the use of different modes with measurement non-invariance may introduce measurement errors that will jeopardize statistical power to detect true differences in self-reported health outcomes, and misinterpret the true change of scores over time.

For all CCSS surveys in both initial and expansions cohorts, the research team sends the survey packages via postal mail to the eligible study participants. For the baseline survey in the CCSS expansion cohort, the survey package includes the questionnaire alongside the link of the web-based surveys. If the participants don't return the survey via either postal mail or on-line, a reminder letter is sent to participants in 6 weeks, followed by a telephone interview for data collection. Although the contents and response options remain the same across the modes of mail, telephone, and web-based surveys, several design factors may threaten the measurement equivalence across these modes.

- Question presentation and response-choice order: By design, CCSS participants read the questions in the mail or web-based surveys or listen to the questions in the telephone interview before they respond to the options. However, research demonstrates that when the questions are presented visually (e.g., mail and web-based surveys), respondents are likely to select the early response options presented if it is agreeable, and then move to the next question. In contrast, when the questions are presented verbally (e.g., telephone survey), respondents are more likely to select the last response options (14,15). Another important design feature of the CCSS is that one or few questions are presented on each individual screen of the web-based survey, whereas more questions are presented on the individual page of the mail survey. Evidence suggests that this feature may produce different answers if the respondents want to refer to previous questions and/or change their response.
- Social desirability: When respondents are involved in social interactions with interviewers, respondents tend to take social norms and expectation into consideration; this is known as social desirability bias in the survey response (14,16). Previous studies have reported that respondents are more likely to provide positive and socially desirable responses in the telephone survey than in the mail or web-based survey. As a result, desirable health status, HRQOL, and lifestyle and behaviors (e.g., regular exercise) tend to be overestimated by the interviewer-administration modes (e.g., face-to-face, telephone interview) compared with the self-administration modes (e.g., mail, web-based survey) (17-19). In contrast, undesirable lifestyle and behaviors (e.g., substance use) and sensitive health problems (e.g., sexual functioning) tend to be underestimated by the interviewer-administration modes compared with the self-administration modes (20). A meta-analysis study comparing social desirability between computer, paper-and-

pencil, and interviewer modes found that there was less social desirability bias in computer administration than in paper-and-pencil and interviewer modes (21).

- Cyber-psychological effect: Individuals with high level of computer anxiety or difficulties in interacting with the computer may report lower scores of health outcomes related to negative mood than those without computer anxiety (22,23). This is especially the case if individuals are older in age and have lower educational background. Interestingly, Dwight et al. compared impression management (a conscious to deceive) and unintentional self-deception (self-deceptive enhancement) between computer mode and either paper-and-pencil or interviewer mode (24). They found that there was less impression bias for computer mode than for non-computer mode though the effect size was small (ES=0.08)

Several empirical studies have compared the equivalence between different modes of health outcome survey. Some studies have reported comparability between paper-and-pencil mode and visual mode (25,26), and between paper-and-pencil, web-based and interactive voice response modes (27-30). A meta-analysis found that the average absolute mean difference was negligible, 1.7% of the score range (1.7 points on a 0-100 scale) between paper-and-pencil mode and computer mode, and 2.4% of the score range between paper-and-pencil mode and personal digital assistant mode (25). However, other studies have revealed systematically lower scores on the electronic mode than on the paper-and-pencil mode (31,32). In cancer research, Smith and colleagues found that the response rate was lower for mail survey than web-based survey on 235 testicular cancer survivors. This study also indicated the percentage of missing data and participant characteristics were comparable between the two modes (33). Zuidgeest and colleagues found that the use of mixed survey (mail survey and web-based survey) can increase the response rate and decrease missing data compared with the use of mail survey alone in 800 breast cancer survivors (34). van den Berg and colleagues conducted a fertility survey in 277 female childhood cancer survivors and found that the use of mixed survey (mail survey and web-based survey) had the same response rate as the use of web-based. However, in the mixed group, respondents with higher educational background were likely to complete the surveys through web-based mode than the mail mode (35).

Although previous studies of mode effects have focused on general population or adult cancer survivors, the generalizability of the results to the childhood cancer survivors is still unclear. This is because childhood cancer survivors who were diagnosed with different cancers (e.g., brain tumor vs. leukemia), have received different treatment exposures (e.g., radio therapy vs. chemotherapy), and are facing various health and life challenges (e.g., visual impairment), may possess different capabilities and preferences in selecting a specific modes of administration. Importantly, the majority of the previous mode effect studies merely compares the response rates between different modes or estimates the correlations of health outcomes collected from different modes. Very few studies have used a rigorous framework to analyze measurement non-invariance in self-reported health outcomes across different modes of survey. Differential item functioning (DIF) analysis is an item-level psychometric method to investigate measurement non-invariance between different modes of administration. This approach explores whether the likelihood of responding to a question of health outcome between different modes is the same or not, while conditioning on the same level of the underlying traits of self-reported outcomes (e.g., anxiety) (36,37). Theoretically, if the underlying anxiety is the same between respondents who are engaged in different survey modes, one should expect that all individuals have the same probability of selecting a response option for a question (e.g., feeling fearful) no matter which survey mode is administered. A DIF exists when this assumption is not held.

Several psychometric methods can be used for DIF analysis (36,37); among different methods, the multiple indicator-multiple cause (MIMIC) method has received greater attention because this method can model item response function and group difference in underlying health outcome simultaneously (38,39). Importantly, the MIMIC method is able to accommodate the background variables (confounding variables) into DIF analysis, which allows a meaningful comparison of health outcome scores among different groups. In the baseline survey of the CCSS expansion cohorts, three modes (mail survey, telephone interview, and web-based survey) were implemented and each mode was comprised of sufficient number of participants for psychometric analyses. As of September 2014, approximately 5,800, 1,500, and 2,300 childhood cancer survivors have completed the surveys through the mail mode, telephone interview, and web-based mode, respectively. Therefore, CCSS provides a great opportunity to test the effects of mode administration.

4. SPECIFIC AIMS/OBJECTIVES/RESEARCH HYPOTHESES

The overall objective of this proposal is to evaluate the effects of administration modes, including mail, telephone, and web-based, using the baseline survey of the CCSS expansion cohort. This proposal will focus on data missingness (i.e., survey completion) and measurement non-invariance of the Brief Symptom Inventory-18 (BSI-18) administered by three modes of the survey.

- Aim 1: To compare the respondent characteristics and data missingness between three modes of survey administration.

  Hypothesis 1a: Respondents who are younger in age and have higher educational attainment are more likely to complete the web-based survey (vs. mail survey or telephone interview) compared with those who are older in age and have lower educational attainment. Respondents who have been diagnosed with brain tumors, have self-reported vision and cognitive impairment, have more severe chronic conditions measured by the Common Terminology Criteria for Adverse Events (CTCAE; e.g., grades 3 and 4) (if available in 2015), have poor self-reported health status (e.g., poor and fair), and use more health care services (e.g., physician visit, hospital admission) are more likely to complete the telephone interview (vs. mail or web-based survey) compared with those who have been diagnosed with cancers other than brain tumors, have no vision and cognitive impairments, have less severe chronic conditions (e.g., CTCAE grades 1 and 2), have good self-reported health status (e.g., excellent and very good), and use less health care services. In addition, respondents who have hearing impairments are more likely to complete mail or web-based survey (vs. telephone interview) compared with those who have no hearing impairments.

  Hypothesis 1b: Responses to questions for the BSI-18, health habits (e.g., smoking, alcohol use, physical activity), and use of health care services (e.g., physician visit, hospital admission) will be less likely to be missing for respondents who are younger in age (<45 years old) compared with those who are younger in age (≥45 years old). Additionally, responses to questions for the BSI-18, health habits, and use of health care services will be less likely to be missing for respondents who participate in the web-based survey and telephone interview compared with those who participate in the mail survey.

- Aim 2: To test measurement non-invariance between three modes of survey administration on each domain of the BSI-18 (depression, anxiety, and somatization) given the same level of underlying distress symptoms with and without controlling for the influence of covariates.

  Hypothesis 2: Respondents who participate in a telephone interview will demonstrate measurement non-invariance in some, but not all, questions of the BSI-18 (i.e., DIF items) compared with those who participate in the mail survey. In contrast, respondents who participate in the web-based survey will demonstrate measurement invariance in the questions of the BSI-18 (i.e., no DIF items) compared with those who participate in the mail survey. Specifically, respondents who participate in the telephone interview will demonstrate more positive response in the DIF items (i.e., reporting less distress symptoms) compared with those who participate in the mail survey given the same level of underlying distress symptoms. The DIF findings reflect the influence of social desirability experienced by respondents in a telephone interview. The same results of measurement non-invariance will also be held if the covariates are taken into account.

- Aim 3: To test psychometric properties of the BSI-18 measured by three modes of survey administration.

  Hypothesis 3: Measurement properties of the BSI-18, including reliability and validity, will be comparable between three modes of survey administration although some questions of the BSI-18 will be identified with DIF. Specifically, reliability of individual BSI-18 domains estimated by Cronbach's alpha coefficients will be comparable between three modes of survey administration. Construct validity of the individual BSI-18 domains estimated by the model fit indices of the confirmatory factor analysis will be comparable between three modes of survey administration. Known-groups validity tested by the scores differences in the individual BSI-18 domains corresponding to different levels of self-reported health status and chronic conditions will be comparable between three modes of survey administration.

- Aim 4: To compare the distress symptoms between respondents who completed three modes of surveys with and without controlling for the influence of DIF items and other covariates.

  Hypothesis 4: Respondents who complete the telephone interview will demonstrate less distress symptoms measured by the SF-18 compared with those who complete the mail or web-based survey due to the social desirability in the survey. However, this result will not hold if the DIF related to mode effects and other covariates, including age, gender, cancer diagnosis, cancer treatment, and health status, are adjusted for in the analyses.

5. METHODS

Subjects:

Adult survivors of childhood cancer in the CCSS expansion cohort who have participated in the baseline survey, and are ≥18 years of age at the time of survey completion.

Outcome of interests:

a. Distress symptoms measured by the BSI-18 (40) (question #K1 through #K18). Percent of subjects with a missing response for each BSI-18 question, compared between three modes of administration. Measurement non-invariance between three modes will also be tested. The BSI-18 is comprised of 18 questions measuring three

domains of distress symptoms, depression, anxiety, and somatization. Because DIF analysis will be conducted at the item level, all questions of the BSI-18 will be used in this study. For each survivor, T-scores on three individual domains and a summary scale (global severity index; GSI) will be generated. The cutoff of 63 on each domains and a summary scale will be used to dichotomize the level of distress symptoms.

b. Health habits:
   Missingness in answering health habit questions between different modes of administration will be tested.
   i. Smoking (questions #O1, #O2, #O3, #O4, #O5, #O6, #O7, and #O8)
   ii. Alcohol use (questions #O9, #O10, #O11, #O12, #O13, and #O14)
   iii. Physical activity (questions #O15, #O16, #O17, and #O20)

Confounding variables:

a. Socio demographic
   i. Age: ≥18 years old.
   ii. Gender: male and female (question #A2).
   iii. Race/ethnicity: White, non-Hispanic; Black, non-Hispanic; Hispanic; and other (questions #A5 and #A5a).
   iv. Education: below high school; high school graduate/ GED; some college/ training after high school; college graduate; postgraduate level; and other (questions #R1 and #R2).
   v. Marital status: married/ living with a partner; widowed/ divorced/ separated; and single (questions #M2 and M3).
   vi. Living arrangement: live with spouse/ partner; live with parents; live with roommate; live with brothers/ sisters; live with other relatives; live alone; and other (question #M1).
   vii. Employment status: working full-time; working part-time; and other (questions #S1 and #S2).
   viii. Insurance status: insured; uninsured; and other (Canadian resident) (questions #U1, #U2, and #U3).
   ix. Incomes: total income of the household (question #T11), people in the household were supported on income (question #T2), and personal income (question #T3).
b. Weight and height: will be used to generate BMI categories – underweight (BMI<18.5 kg/m2); normal weight (BMI: 18.5 – 24.9 kg/m2); overweight (BMI: 25.0 – 29.9 kg/m2); and obese (BMI: ≥30 kg/m2) (questions #A3 and #A4).
c. Internet use per week (question #A10).
d. Cancer diagnosis
   i. Primary cancer: leukemia; central nervous system (CNS) tumor; Hodgkin lymphoma; Non-Hodgkin lymphoma; Wilms tumor; neuroblastoma; soft tissue sarcoma; bone tumor; and other.
   ii. Second cancer or cancer recurrence (separate data source: SMN validated data base).
e. Survival time: will be generated using the following two variables
   i. Age at diagnosis: in years.
   ii. Age at interview: in years.
f. Cancer treatment
   i. Chemotherapy: none; methotrexate; corticosteroid; anthracyclines; alkylating agents; and other chemotherapy (all with yes/ no).

ii. Radiotherapy: none; cranial radiotherapy; and other radiotherapy (all with yes/ no).
iii. Surgery: none; amputation; and other surgery (all with yes/ no).

g. Health status and psychological outcomes
   i. Self-reported health outcomes
      1. General health status (question #O21).
      2. Health status transition (question #O22).
      3. Hearing impairment (questions #C1, #C2, #C3, and #C6).
      4. Vision impairment (questions #C8 and #C9).
   ii. Type and severity of chronic conditions:
      1. Type of chronic conditions: major joint replacement; congestive heart failure; second malignant neoplasm; cognitive dysfunction, severe; coronary artery disease; cerebrovascular accident; renal failure or dialysis; hearing loss not corrected by aid; legally blind or loss of an eye; ovarian failure; and other.
      2. Grading of chronic conditions based on CTCAE (if data are available in 2015): mild (Grade 1); moderate (Grade 2); severe (Grade 3); or life-threatening or disabling (Grade 4). Each individual may possess a variety of chronic conditions with different grades. However, the high grade will be used to represent the severity of chronic conditions of an individual.

h. Use of health care services
   i. Physician visit: question #B1, #B2, #B3, #B4, and #B5.
   ii. Hospital admission: question #B6.

Analytic approach**:**

For Aim 1: To compare the respondent characteristics and data missingness between three modes of survey administration.

Distributions of respondents' characteristics will be reported by each mode of survey administration (Table 1a). For continuous variables (e.g., age), mean and standard deviation will be reported; ANOVA tests will be conducted to compare the difference between three modes of administration. For categorical variables (e.g., race/ethnicity), count and percentage will be reported; chi-square tests will be conducted to compare the difference between three modes of survey administration.

Percent of subjects with a missed response to individual questions ("missingness") of the BSI-18 will be reported by three modes, respectively, and by two age strata, respectively (Table 1b). Chi-square tests will be conducted to compare the percent of missing responses in individual questions between three modes of administration. Similarly, chi-square tests will be conducted to compare the percent of missing responses in individual questions between two age strata.

For Aim 2: To test measurement non-invariance between three modes of survey administration on each domain of the BSI-18 (depression, anxiety, and somatization) given the same level of underlying distress symptoms with and without controlling for covariates.

The analyses for measurement non-invariance will be conducted using the structural equation modeling framework with the estimator of the weighted least square with mean and variance-adjusted (WLSMV). WLSMV does not assume normally distributed data

and is a robust estimator for handling items with categorical response options. WLSMV estimator performs especially well for the data with ceiling and floor effect (41). In this study, two sets of MIMIC models will be conducted for testing DIF related to mode of administration: the first set includes three modes of administration (as the main effect of DIF) and questions of the individual BSI-18 domains; the second set includes three modes of administration, BSI-18 questions, and other covariates.

In MIMIC, a three-step analytic scheme will be used to identify DIF items related to model of administration (42). First, DIF-free items will be identified and the remained items will be treated as studied items. Second, each studied item will be tested individually. Third, a final model will be constructed to accommodate DIF items related to modes of administration. Estimates of discrimination parameter, threshold parameter, mean difference on the BSI-18 by modes of administration, and DIF effects from the final model (Table 3). The specific operational procedures are described as follows:

Step 1: Preliminary analyses will be performed to select a subset of DIF-free items. Each item will be tested for DIF with all other items presumed DIF-free. Practically, one item at a time will be regressed on the variable of mode administration (Figure 1). A specific item j will be regarded as DIF-free if the discrimination parameter (α) is at least 0.5 and the regression coefficient (β) indicating the difference in the item threshold for the item j between the modes of administration (e.g., web-based vs. paper-and-pencil) is not significant ($p>0.05$).

Step 2: Items that are not assigned to DIF-free subset (i.e., studied items for potential DIF) will be tested individually using the DIFTEST procedure for the nested models (i.e., full and constrained models with different constraints on item parameters). Instead of comparing the chi-square differences between the nested models, the use of the DIFTEST procedure is more appropriate for the WLSMV estimators that deal with the categorical items of the BSI-18. To test studied item j for DIF, a full model will compared with a more constrained model. In both models, items assigned to DIF-free subset will not be regressed on the variable of mode administration. Specifically, in the full model, all studied items are regressed on the variable of mode administration (i.e., regression coefficients are feely estimated or regression coefficients $\neq 0$); in the constrained model, item j will not be regressed on the variable of mode administration (i.e., invariant on the regression coefficient or regression coefficients = 0). A significant difference between the nested models (full and constrained models) will suggest that the good model fit will significantly declines if the item j is assumed to be DIF-free (i.e., regression coefficients = 0). As a result, the item j is a DIF item.

Step 3: A final MIMIC mode will be constructed for individual domain of the BSI-18 (Table 2). In the final model, only items that show significant DIF will be regressed on the variable of mode administration. Additionally, the domains of the BSI-18 will be regressed on the variable of mode administration. Item discrimination parameter (α), item threshold parameter (ζ), mean difference on the individual BSI-18 domain scores related to mode administration (γ), and DIF of individual items (β) will be estimated (Figure 1). A negative value of β will indicate that the item response is smaller on one mode of administration (i.e., telephone interview or web-based mode) than the paper-and-penile mode (the reference group) given the same level of underlying distress symptoms.

For Aim 3: To test psychometric properties of the BSI-18 measured by three modes of survey administration.

For evaluating scale reliability, Cronbach's alpha coefficients will be estimated for individual domains of the BSI-18 by three modes of survey administration (Table 3a). If the alpha value ≥0.7 is observed across three modes of survey administration, scale reliability of the BSI-18 between three modes is acceptable (43).

For evaluating construct validity, a confirmatory factor analysis will be performed to examine the extent to which the construct of the individual BSI-18 domains is comparable between three modes of survey administration (Table 3b). Specifically, factorial structure of the BSI-18 data collected from three modes of survey will be tested. For each mode, two indicators will be estimated to determine the goodness of model fit to the data: the comparative fit index (CFI) and root mean square error of approximation (RMSEA). If the values of CFI >0.9 and RMSEA <0.06 are determined across three modes of survey administration, construct validity of the BSI-18 between three modes is comparable (44).

Known-groups validity will be assessed to examine the extent to which the individual domains of the BSI-18 discriminate between respondents who have different levels of chronic conditions measured by CTCAE and health status (i.e., self-reported health status) (Table 3c). To facilitate comparisons, CTCAE will be categorized by two groups that capture less severe chronic conditions (grades 1 and 2) and more severe chronic conditions (grade 3 and 4); self-reported health status will be categorized by two groups that capture low health status (poor, fair) and high health status (good, very good, excellent). For individual mode, t-tests will be performed to examine the difference in the BSI-18 domains scores by different levels of health status. Next, values of the relative validity will be estimated for examining the known-groups validity across three modes of survey administration (45). Relative validity is defined as the F-value (square of the t-value) on one mode divided by the F-value on the reference mode (a mode with the lowest t-value). Comparable known-groups validity will be determined if the values of relative validity are below 1 across three modes.

For Aim 4: To compare the distress symptoms between respondents who completed three modes of surveys with and without controlling for the influence of DIF items and other covariates.

Linear regression analyses will be conducted to test the differences in the underlying domain scores of the BSI-18 between three modes of survey administrations with and without adjusting for DIF effects (Table 4). This procedure will suggest the impact of measurement non-invariance in the BSI-18 related to the mode effects. Adjusting DIF in the analysis will be performed through freely estimating the coefficients β for the mode administration associated with individual BSI-18 questions that are identified with DIF (a dotted line in Figure 1). Additionally, the impact of DIF on the change of BSI-18 domain scores for individuals will be evaluated with and without adjusting for DIF effects. The changes in the BSI-18 scores ≥0.2 unit of effect size will be deemed as the evidence of minimally important change (46).

Software**:**

For Aims 1 and 3, the analyses will be conducted using STATA 13; for Aims 2 and 4, the analyses will be conducted using Mplus 7.3. All analysis will be conducted by PI: I-Chan

Huang at the St. Jude Children's Research Hospital, with review of results and manuscript carried out by members of the CCSS Statistical Center.

References:

1. Jemal A, Siegel R, Ward E, Hao Y, Xu J, Thun MJ. Cancer statistics, 2009. *CA Cancer J Clin*. 2009;59(4):225-249.

2. Robison LL, Hudson MM. Survivors of childhood and adolescent cancer: Life-long risks and responsibilities. *Nat Rev Cancer*. 2014;14(1):61-70.

3. Mariotto AB, Rowland JH, Yabroff KR, et al. Long-term survivors of childhood cancers in the united states. *Cancer Epidemiol Biomarkers Prev*. 2009;18(4):1033-1040.

4. Oeffinger KC, Mertens AC, Sklar CA, et al. Chronic health conditions in adult survivors of childhood cancer. *N Engl J Med*. 2006;355(15):1572-1582.

5. Armstrong GT, Kawashima T, Leisenring W, et al. Aging and risk of severe, disabling, life-threatening, and fatal events in the childhood cancer survivor study. *J Clin Oncol*. 2014;32(12):1218-1227.

6. Hudson MM, Mertens AC, Yasui Y, et al. Health status of adult long-term survivors of childhood cancer: A report from the childhood cancer survivor study. *JAMA*. 2003;290(12):1583-1592.

7. Zeltzer LK, Recklitis C, Buchbinder D, et al. Psychological status in childhood cancer survivors: A report from the childhood cancer survivor study. *J Clin Oncol*. 2009;27(14):2396-2404.

8. Aaronson NK, Ahmedzai S, Bergman B, et al. The european organization for research and treatment of cancer QLQ-C30: A quality-of-life instrument for use in international clinical trials in oncology. *J Natl Cancer Inst*. 1993;85(5):365-376.

9. Cella D, Yount S, Rothrock N, et al. The patient-reported outcomes measurement information system (PROMIS): Progress of an NIH roadmap cooperative group during its first two years. *Med Care*. 2007;45(5 Suppl 1):S3-S11.

10. Bloom DE. Technology, experimentation, and the quality of survey data. *Science*. 1998;280(5365):847-848.

11. Tiplady B. Electronic patient diaries and questionnaires - ePRO now delivering on the promise? *Patient*. 2010;3(3):179-183.

12. 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):419-429.

13. Eremenco S, Coons SJ, Paty J, et al. PRO data collection in clinical trials using mixed modes: Report of the ISPOR PRO mixed modes good research practices task force. *Value Health*. 2014;17(5):501-516.

14. Tourangeau R, Smith TW. Asking sensitive questions the impact of data collection mode, question format, and question context. *Public Opin Q*. 1996;60(2):275-304.

15. Bowling A. Mode of questionnaire administration can have serious effects on data quality. *J Public Health (Oxf)*. 2005;27(3):281-291.

16. Presser S, Stinson L. Data collection mode and social desirability bias in self-reported religious attendance. *Am Sociol Rev*. 1998;61(1):137-145.

17. Lyons RA, Wareham K, Lucas M, Price D, Williams J, Hutchings HA. SF-36 scores vary by method of administration: Implications for study design. *J Public Health Med*. 1999;21(1):41-45.

18. Perkins JJ, Sanson-Fisher RW. An examination of self- and telephone-administered modes of administration for the australian SF-36. *J Clin Epidemiol*. 1998;51(11):969-973.

19. Weinberger M, Oddone EZ, Samsa GP, Landsman PB. Are health-related quality-of-life measures affected by the mode of administration? *J Clin Epidemiol*. 1996;49(2):135-140.

20. Aquilinom WS, LoSciuto LA. Effects of interview mode on self-reported drug use. *Public Opin Q*. 1990;54(3):362-393.

21. Richman WL, Kiesler S, Weisband S, Drasgow F. A meta-analytic study of social desirability distortion in computer-administered questionnaires, traditional questionnaires, and interviews. *J Appl Psychol*. 1999;84(5):754-775.

22. 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(2):315-321.

23. 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 (Pt 4):599-610.

24. Dwight SA. A quantitative review of the effect of computerized testing on the measurement of social desirability. *Educ Psychol Meas*. 2000;60(3):340-360.

25. Gwaltney CJ, Shields AL, Shiffman S. Equivalence of electronic and paper-and-pencil administration of patient-reported outcome measures: A meta-analytic review. *Value Health*. 2008;11(2):322-333.

26. Ramachandran S, Lundy JJ, Coons SJ. Testing the measurement equivalence of paper and touch-screen versions of the EQ-5D visual analog scale (EQ VAS). *Qual Life Res*. 2008;17(8):1117-1120.

27. Athale N, Sturley A, Skoczen S, Kavanaugh A, Lenert L. A web-compatible instrument for measuring self-reported disease activity in arthritis. *J Rheumatol*. 2004;31(2):223-228.

28. McCabe SE, Diez A, Boyd CJ, Nelson TF, Weitzman ER. Comparing web and mail responses in a mixed mode survey in college alcohol use research. *Addict Behav*. 2006;31(9):1619-1627.

29. Lundy JJ, Coons SJ. Measurement equivalence of interactive voice response and paper versions of the EQ-5D in a cancer patient sample. *Value Health*. 2011;14(6):867-871.

30. Lundy JJ, Coons SJ, Aaronson NK. Testing the measurement equivalence of paper and interactive voice response system versions of the EORTC QLQ-C30. *Qual Life Res*. 2014;23(1):229-237.

31. Bent H, Ratzlaff CR, Goligher EC, Kopec JA, Gillies JH. Computer-administered bath ankylosing spondylitis and quebec scale outcome questionnaires for low back pain: Agreement with traditional paper format. *J Rheumatol*. 2005;32(4):669-672.

32. Bellamy N, Campbell J, Stevens J, Pilch L, Stewart C, Mahmood Z. Validation study of a computerized version of the western Ontario and McMaster universities VA3.0 osteoarthritis index. *J Rheumatol*. 1997;24(12):2413-2415.

33. Smith AB, King M, Butow P, Olver I. A comparison of data quality and practicality of online versus postal questionnaires in a sample of testicular cancer survivors. *Psychooncology*. 2013;22(1):233-237.

34. Zuidgeest M, Hendriks M, Koopman L, Spreeuwenberg P, Rademakers J. A comparison of a postal survey and mixed-mode survey using a questionnaire on patients' experiences with breast care. *J Med Internet Res*. 2011;13(3):e68.

35. van den Berg MH, Overbeek A, van der Pal HJ, et al. Using web-based and paper-based questionnaires for collecting data on fertility issues among female childhood cancer survivors: Differences in response characteristics. *J Med Internet Res*. 2011;13(3):e76.

36. Teresi JA. Different approaches to differential item functioning in health applications. advantages, disadvantages and some neglected topics. *Med Care*. 2006;44(11 Suppl 3):S152-70.

37. Teresi JA, Fleishman JA. Differential item functioning and health assessment. *Qual Life Res*. 2007;(16 Suppl 1):33-42.

38. Jones RN. Identification of measurement differences between English and Spanish language versions of the mini-mental state examination. detecting differential item functioning using MIMIC modeling. *Med Care*. 2006;44(11 Suppl 3):S124-33.

39. Carle AC. Mitigating systematic measurement error in comparative effectiveness research in heterogeneous populations. *Med Care*. 2010;48(6 Suppl):S68-74.

40. Recklitis CJ, Parsons SK, Shih MC, Mertens A, Robison LL, Zeltzer L. Factor structure of the brief symptom inventory--18 in adult survivors of childhood cancer: Results from the childhood cancer survivor study. *Psychol Assess*. 2006;18(1):22-32.

41. Brown TA. *Confirmatory factor analysis for applied research.* New York: Guilford Press; 2006.

42. Woods CM, Oltmanns TF, Turkheimer E. Illustration of MIMIC-model DIF testing with the schedule for nonadaptive and adaptive personality. *J Psychopathol Behav Assess*. 2009;31(4):320-330.

43. Fayers PM, Machin D. *Quality of life : Assessment, analysis, and interpretation.* Chichester; New York: John Wiley; 2000.

44. Hu L, Bentler BM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Struct Equ Modeling*. 1999;6(1):1-55.

45. McHorney CA, Ware JE,Jr, Raczek AE. The MOS 36-item short-form health survey (SF-36): II. psychometric and clinical tests of validity in measuring physical and mental health constructs. *Med Care*. 1993;31(3):247-263.

46. Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: The remarkable universality of half a standard deviation. *Med Care*. 2003;41(5):582-592.

APPENDIX: SAMPLE TABLES AND FIGURES

Table 1a: Characteristics of study population (N= 2,377)

	Mail survey (N = 1,452)	Telephone interview (N = 392)	Web-based survey (N = 533)	p-value
Age, (mean, SD) [in year]				
Sex, (N, %)				
Male				
Female				
Race/ethnicity, (N, %)				
White, non-Hispanic				
Black, non-Hispanic				
Hispanic				
Other				
Educational background, (N, %)				
Below high school				
High school graduate/ GED				
Some college/ training after high school				
College graduate				
Post graduate level				
Marital status, (N, %)				
Married/ living with a partner				
Widowed/ divorced/ separated				
Single				
Employment status, (N, %)				
Working full-time				
Working part-time				
Others (will breakdown depending upon frequency)				
Insurance status, (N, %)				
Insured				
Uninsured				
Annual household incomes, (N, %)				
< $19,999				
$20,000 – $39,999				
$40,000 – $59,999				
$60,000 – $79,999				
$80,000 – $99,999				
≥ $100,000				
Age at diagnosis, (mean, SD) [in year]				
Age at interview, (mean, SD) [in year]				
Years since diagnosis, (mean, SD)				

Cancer diagnosis, (N, %)				
Leukemia				
Central nervous system (CNS) tumor				
Hodgkin lymphoma				
Non-Hodgkin lymphoma				
Wilms tumor				
Neuroplastoma				
Soft tissue sarcoma				
Bone tumor				
Second cancer, (N, %)				
Yes				
No				
Chemotherapy, (N, %)				
None				
Methotrexate				
Corticosteroid				
Anthracyclines				
Alkylating agents				
Other				
Radiotherapy, (N, %)				
None				
Cranial				
Other				
Surgery, (N, %)				
None				
Amputation				
Other				
Severity of chronic condition by CTCAE, (N, %)[†]				
Grade 1				
Grade 2				
Grade 3				
Grade 4				
Self-reported general health status, (N, %)				
Excellent				
Very good				
Good				
Fair				
Poor				
Body mass index (BMI), (N, %)				
Underweight (<18.5 kg/m^2)				
Normal weight (18.5 – 24.9 kg/m^2)				
Overweight (25.0 – 29.9 kg/m^2)				
Obese (≥30 kg/m^2)				
Brief Symptom Inventory, (N, %)				
Anxiety (cutoff: 63)				

Depression (cutoff: 63)				
Somatization (cutoff: 63)				
Global severity index (cutoff: 63)				
Health habit, (N, %)				
Currently use cigarette				
Currently chew tobacco				
Currently use snuff tobacco				
Currently use pipes				
Currently use cigars				
Had alcohol drink during the last 12 months				
Did exercise or sports in the last 7 days for ≥20 minutes				
Use of health care services, (N, %)				
Physician visit (yes/no)				
Hospital admission (yes/no)				

† If CTCAE data are available in 2015

Table 1b: Missingness of the BSI-18, health habit questions, and use of health care services between three modes of survey administration and age strata[†]

	Mode				Age		
	Mail survey (N = 1,452)	Telephone interview (N = 392)	Web-based survey (N = 533)	p-value	<45 years old (N =)	≥45 years old (N =)	p-value
BSI-18							
Item 01							
Item 02							
Item 03							
Item 04							
Item 05							
Item 06							
Item 07							
Item 08							
Item 09							
Item 10							
Item 11							
Item 12							
Item 13							
Item 14							
Item 15							
Item 16							
Item 17							
Item 18							
Depression score							
Anxiety score							
Somati-zation score							
Global severity index							
Health habit							
Cigarettes							
Chewing tobacco							
Snuff tobacco							
Pipes							
Cigars							
Drinks containing alcohol							

Exercise or sports for ≥20 minutes							
Use of health services							
Physician visit							
Hospital admission							

† % of subjects in each column missing response for the individual questions will be reported in the Table

Table 2: DIF in BSI-18 items related to modes of survey administration (the final model)†

	Modes	Difference in item score (γ)	X^2 (p-value)	Item parameters (standard error)† from the final model				
				Factor loading (α)	1st threshold (ζ_1)	2nd Threshold (ζ_2)	3rd Threshold (ζ_3)	4th Threshold (ζ_4)
Depression domain								
Item 1	Phone vs. Mail							
	Web vs. Mail							
Item 2	Phone vs. Mail							
	Web vs. Mail							
Item 3	Phone vs. Mail							
	Web vs. Mail							
Anxiety domain								
Item 1	Phone vs. Mail							
	Web vs. Mail							
Item 2	Phone vs. Mail							
	Web vs. Mail							
Item 3	Phone vs. Mail							
	Web vs. Mail							
Somatization domain								
Item 1	Phone vs. Mail							
	Web vs. Mail							
Item 2	Phone vs. Mail							
	Web vs. Mail							
Item 3	Phone vs. Mail							
	Web vs. Mail							

† If an item is identified with DIF, different values of a parameter will be reported by different modes.

Table 3a: Reliability of individual BSI-18 domains between three modes of survey administration†

	Mail survey	Telephone interview	Web-based survey
Depression			
Anxiety			
Somatization			
Global severity index			

† Coefficient of Cronbach's alpha will be reported in the table.

Table 3b: Construct validity of individual BSI-18 domains between three modes of survey administration

	Mail survey	Telephone interview	Web-based survey
Depression X^2 RMSEA CFI			
Anxiety X^2 RMSEA CFI			
Somatization X^2 RMSEA CFI			
Global severity index X^2 RMSEA CFI			

Table 3c: Known-groups validity of individual BSI-18 domains between three modes of survey administration

	Mail survey	Telephone interview	Web-based survey
Depression			
CTCAE Grades 3, 4 vs. 1, 2†			
Score difference			
Relative validity			
Health status poor, fair vs. good, very good, excellent			
Score difference			
Relative validity			
Anxiety			
CTCAE Grades 3, 4 vs. 1, 2†			
Score difference			
Relative validity			
Health status poor, fair vs. good, very good, excellent			
Score difference			
Relative validity			
Somatization			
CTCAE Grades 3, 4 vs. 1, 2†			
Score difference			
Relative validity			
Health status poor, fair vs. good, very good, excellent			
Score difference			
Relative validity			
Global severity index			
CTCAE Grades 3, 4 vs. 1, 2†			
Score difference			
Relative validity			
Health status poor, fair vs. good, very good, excellent			
Score difference			
Relative validity			

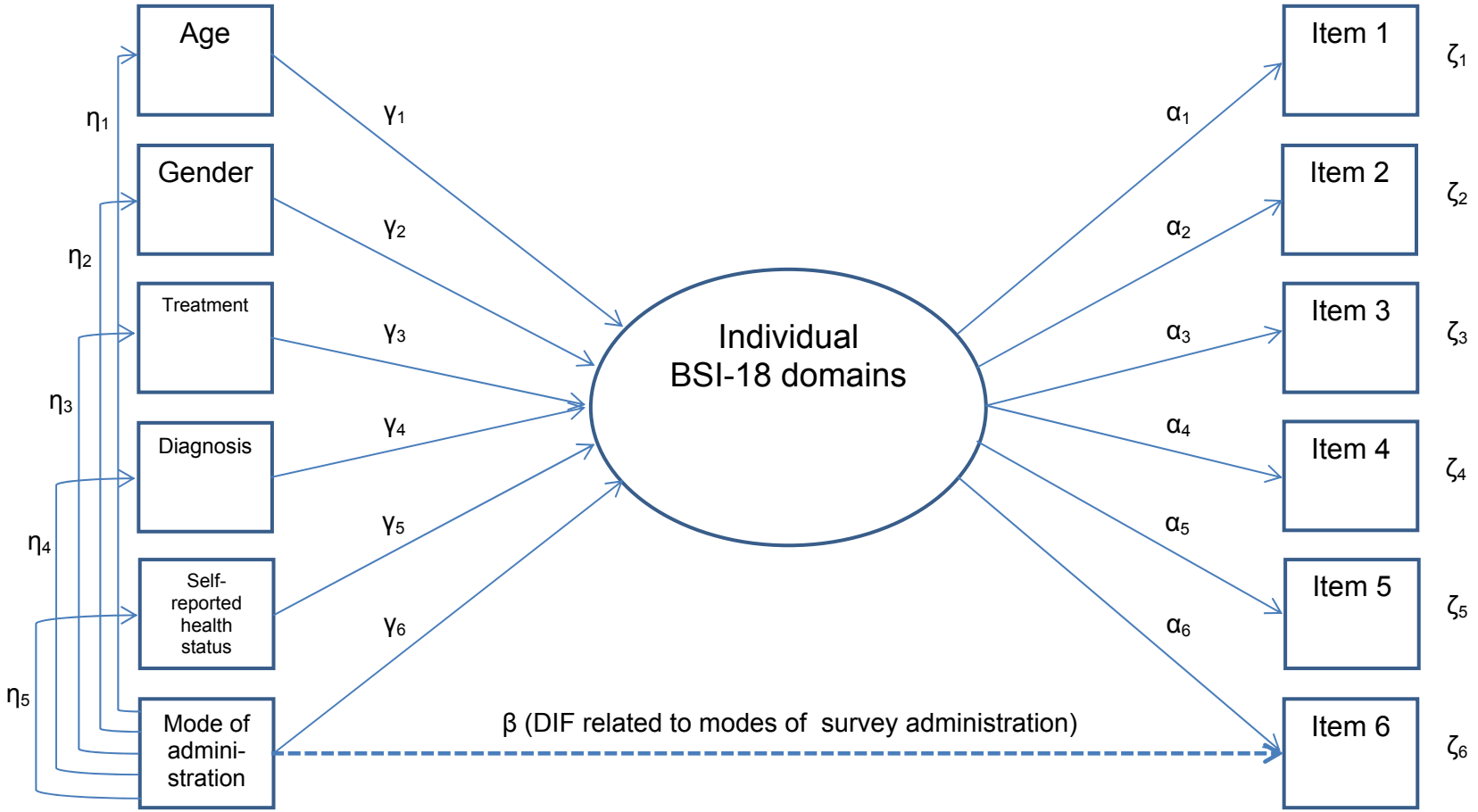
† If CTCAE data are available in 2015

Table 4: Distress symptom scores between three modes of survey administration before and after DIF adjustment[†]

	Before DIF adjustment	After DIF adjustment	Respondents who increase or decrease scores by ≥2 SD, respectively, after adjustment (%)
Depression			
Mail survey			
Telephone interview			
Web-based survey			
p-value			
Anxiety			
Mail survey			
Telephone interview			
Web-based survey			
p-value			
Somatization			
Mail survey			
Telephone interview			
Web-based survey			
p-value			
Global severity index			
Mail survey			
Telephone interview			
Web-based survey			
p-value			

[†] Will adjust for important covariates including age, gender, education, race/ethnicity, type of cancer, cancer therapy, chronic conditions by CTCAE (if data were available in 2015), and self-reported health status

Figure 1: A conceptual model for testing DIF related to the modes of administration



η: correlations between covariates; γ: associations of covariates with individual BSI-18 domains; β: associations of the mode administration and individual items conditioning on BSI-18 domain score (i.e., DIF effect); α: item factor loading; ζ: item thresholds.