The Surgical Education Culture Optimization through Targeted Interventions Based on National Comparative Data (“SECOND”) Trial

STUDY PROTOCOL

DRAFT: 11 November 2021

SECOND TRIAL STUDY PROTOCOL

CONTENTS

SECOND TRIAL REGISTRATION INFORMATION

The **S**urgical **E**ducation **C**ulture **O**ptimization through Targeted Interventions Based on **N**ational Comparative **D**ata ("SECOND") Trial is registered at NCT03739723 at www.clinicaltrials.gov.

The SECOND Trial website is located at www.thesecondtrial.org.

SECOND TRIAL INVESTIGATORS AND STUDY TEAM

Principal Investigators

Karl Bilimoria, MD, MS
Northwestern University Feinberg School of Medicine

Yue-Yung Hu, MD, MS
Northwestern University Feinberg School of Medicine

Investigators

David Hoyt, MD
American College of Surgeons

Thomas Nasca, MD
Accreditation Council for Graduate Medical Education

Project Coordinator

Daniela Amórtegui, MA
Northwestern University Feinberg School of Medicine

Additional Study Team Members

Gaurava Agarwal, MD
Northwestern University Feinberg School of Medicine

Elaine Cheung, PhD
Northwestern University Feinberg School of Medicine

Jeanette Chung, PhD
Northwestern University Feinberg School of Medicine

Joshua Eng, PhD
Northwestern University Feinberg School of Medicine

Caryn Etkin, PhD
Northwestern University Feinberg School of Medicine

Julie Johnson, PhD
Northwestern University Feinberg School of Medicine

Larry Hedges, PhD
Northwestern University Feinberg School of Medicine

Reiping Huang, PhD
Northwestern University Feinberg School of Medicine

SECOND TRIAL OVERSIGHT

Steering Committee

Jo Buyske, MD,
Executive Director, American Board of Surgery

David Hoyt, MD
Executive Director, American College of Surgeons

Thomas Nasca, MD
President and CEO, Accreditation Council for Graduate Medical Education

Faculty Advisory Committee

Chandrakanth Are, MD, MBA
Designated Institutional Official, University of Nebraska

Jennifer Choi, MD
Program Director, General Surgery, Indiana University

Joshua Goldstein, MD
Designated Institutional Official, Northwestern University

Caprice Greenberg, MD, MPH
Professor of Surgery, Mortgage Distinguished Chair in Health Services Research, University of Wisconsin

Jacob Greenberg, MD, MEd
Program Director, General Surgery, University of Wisconsin

Sean Grondin, MD, MPH
Chair, University of Calgary

Benjamin Jarmin, MD
Program Director, General Surgery, Gundersen Health System

Anne Mosenthal, MD
Chair, Department of Surgery, Rutgers University

Todd Nickloes, DO, MBA
President, American College of Osteopathic Surgeons

Michael Nussbaum, MD
Chair, Department of Surgery, Virginia Tech Carilion School of Medicine

Taylor Riall, MD, PhD
Chair, Department of Surgery, University of Arizona

Kari Rosenkranz, MD

Program Director, Dartmouth-Hitchcock Medical Center

Douglas Smink, MD, MPH
Program Director, General Surgery, Brigham & Women's Hospital

Patricia Turner, MD
Director, Member Services, American College of Surgeons

Resident Advisory Committee

Christie Buonpane, MD
Resident, Geisinger Clinic

Sara Daniel, MD
Resident, University of Washington

Woo Do, MD
Madigan Army Medical Center

Ryan Ellis, MD
Resident, Northwestern University

Micaela Esquivel, MD
Resident, Stanford University

Charity Glass, MD
Resident, Beth Israel Deaconess Medical Center

Farnaz Haji, MD
Resident, Larkin Hospital

Rebecca Hoffman, MD
Chair, Resident and Associate Society, American College of Surgeons
Fellow, Washington University

Meixi Ma, MD
Resident, University of Alabama

Ryland Stucke, MD
Resident, Dartmouth-Hitchcock Medical Center

Lindsey Zhang, MD
Resident, University of Chicago

Bioethics Panel

Peter Angelos, MD, PhD
Associate Director, MacLean Center for Clinical Medical Ethics, University of Chicago

Allen Peetz, MD
Vanderbilt University Medical Center

Jennifer Tseng, MD
University of Chicago

Christian Vercler, MD
University of Michigan

Anji Wall, MD, PhD
Baylor University Medical Center

SECTION A

INTRODUCTION

Resident well-being remains a persistent, concerning issue. Throughout the course of residency, trainees experience significant stressors, including heavy workloads, adverse patient events, an increasingly complex body of knowledge to master, and limited time with family and support structures. Subject to a power differential, they also are susceptible to mistreatment, such as discrimination, harassment, and abuse.[1, 2] This combination of stressors and/or mistreatment results in potentially "toxic" learning environments that negatively impact well-being, leading to adverse outcomes such as burnout,[3, 4] attrition,[5, 6] and suicide.[2, 3, 7]

Through a six-year partnership with the Accreditation Council Graduate Medical Education (ACGME), the American College of Surgeons (ACS), the American Board of Surgery (ABS), and the Association of Program Directors in Surgery (APDS), our team created the largest consortium of general surgery programs in the country (N=121) in order to conduct the Flexibility In duty hour Requirements for Surgical Trainees (FIRST) Trial (www.TheFirstTrial.org).[8-14] This work spurred national policy change for residents across all specialties through the ACGME 2017 Common Program Requirements revisions.[15] As part of FIRST Trial follow-up, we administer an annual survey in conjunction with the ABS In-Training Examination (ABSITE), which is taken by all residents training in U.S. ACGME-accredited general surgery programs. This survey achieves an unparalleled 99% response rate, thus avoiding the non-response bias of nearly every prior study. Thus, we were uniquely able to quantify surgical resident well-being: 39% of all surgical residents experience weekly burnout symptoms, 12.6% think of leaving their programs,[16] and 4.5% admit to suicidal thoughts.[2]

To further understand the survey findings, our research team conducted annual interviews and focus groups with residents, faculty, and program directors across the country. This qualitative work highlighted two critical issues impeding programs' abilities to address issues with their learning environments and/or their residents' well-being: (1) Programs are unable to self-assess; although they recognize the importance of the learning environment and their residents' well-being, they lack the means to objectively measure their performance in these domains. Standard instruments exist for only a narrow range of relevant environmental exposures and/or well-being outcomes. Programs with the initiative and resources to identify and administer appropriate instruments subsequently struggle with interpretation. Published data exist only for small subsets of residents (e.g., a handful of residency programs), and it is unclear if they are representative of the entire cohort (i.e., all general surgery residents). Moreover, different study groups analyze existing instruments differently,[17] precluding comparison. (2) Even for the subset of programs that are able to pinpoint specific problems in their learning environments and/or residents' well-being, there is little to guide them in how to intervene and improve. Few interventions exist, and there is little robust data to support their adoption, particularly outside of the home institutions in which they were developed. Moreover, programs must seek interventions out, generally through word of mouth, as there is no standardized format for sharing with others.

Finally, our qualitative work with residents, faculty, and program directors revealed concerns about the potential negative impact of wellness interventions on training (e.g., more time off may result in less clinical experience). Thus, there is a need to assess the impact of wellness interventions on resident education and training.

SECTION B

STUDY OBJECTIVES

To transform the learning environment in general surgery residency, we will again partner with the ACGME, ACS, and ABS, along with the Association of Program Directors in Surgery (APDS), Society of Surgical Chairs (SSC), the Association for American Medical Colleges (AAMC), and the Residents and Associates Society of the ACS (RAS-ACS) to conduct the national Surgical Education Culture Optimization through targeted interventions based on National comparative Data (SECOND) Trial. The SECOND Trial is another national, pragmatic, cluster-randomized trial in which enrolled surgical residency programs will be randomized to either a "Control" or a multi-faceted "Intervention" condition.

Each program will receive nationally benchmarked data on residents' well-being through three specific outcomes: burnout, thoughts of attrition, and suicidal thoughts. "Intervention" programs will additionally receive (1) a Learning Environment Report, containing more granular data (e.g., duty hour violations, autonomy, sexual harassment), (2) a Wellness Toolkit of ready-to-implement interventions, and (3) implementation support, consisting of process improvement coaching, peer mentoring, collaboration networking, and topical experts to assist programs in implementing Toolkit solutions, adapted to their local contexts.

Objective 1: To develop a comprehensive, multifaceted Wellness Toolkit of potential interventions to improve the resident learning environment.
Expected Outcome: We will undertake four approaches to develop a comprehensive, multifaceted Wellness Toolkit of ready-to-implement interventions: (1) a thorough review of literature and best practice guidelines, (2) a survey-based inventory of wellness initiatives at each program, (3) phone interviews with program directors to gather further information about initiatives reported in the inventory, and, most importantly, (4) two-day Program Tours of representative programs that will provide insight about the various factors (e.g., environmental, infrastructural, and cultural) that impact the learning environment and resident well-being.

Objective 2: To assess the impact of the SECOND Trial Intervention (Learning Environment Reports, Wellness Toolkit, implementation support) on (A) resident well-being, (B) resident education, and (C) patient outcomes.
Expected Outcome: A pragmatic cluster-randomized trial will demonstrate that access to personalized, nationally-benchmarked data, along with an array of ready-to-implement interventions and implementation assistance, will (A) measurably improve the learning environment and resident well-being without (B) negatively impacting resident education or (C) patient care.

Objective 3: To refine the Wellness Toolkit by identifying successful context-specific interventions and expanding intervention access to the Control arm.
Expected Outcome: By examining specific initiatives implemented by programs as well as their successes and failures, we will be able to identify (1) particular interventions, (2) groups of interventions, or (3) configurations of interventions that are beneficial in certain contexts (e.g., large, urban programs and small, rural programs have unique needs requiring different interventions). Based on these data, the Wellness Toolkit will be refined, and access will be expanded to all participating residency programs.

SECTION C

STUDY ENDPOINTS

C.1. PRIMARY ENDPOINT

The primary study endpoint is resident burnout, as measured by the abbreviated Maslach Burnout Inventory[2, 3, 17-19] , which is incorporated into the SECOND Trial Resident Survey. Power calculations were based upon this endpoint. Details on measurement are in §E1.

C.2. SECONDARY ENDPOINTS

C.2.1 Resident Well-Being: Our conceptual model is adapted from Shanafelt's Areas of Work-Life,[20] which posits that organizational performance on various domains impacts clinician well-being. These domains (and their component items) will serve as secondary endpoints in the measurement of resident well-being:

- Workload & Job Demands
- Efficiency & Resources
- Organizational Culture
- Control & Flexibility
- Meaning in Work
- Resident Camaraderie
- Faculty Engagement
- Work-Life Integration

Based upon our 2018 ABSITE Survey data,[2, 21] we also added the following:

- Mistreatment

In 2019, we added a series of questions to address these domains. Confirmatory Factor Analysis was performed on the responses to validate these metrics and their fit within our conceptual model. See **Appendix 1** for Confirmatory Factor Analysis.

C.2.2 Resident Education: The following secondary endpoints will be examined to assess the impact of wellness interventions on resident education:

- ABSITE scores
- ABS Qualifying Examination (QE) scores
- ABS Certifying Examination (CE) scores
- SECOND Trial Resident Survey
- ACGME Case Log data
- ACGME Annual Resident and Faculty Surveys

C.2.3 Patient Outcomes: The following secondary endpoints will be examined to assess the impact of wellness interventions on patient outcomes in the subset of programs that participate in the American College of Surgeons National Surgical Quality Improvement Program (ACS NSQIP):[8, 22-24]

- 30 day postoperative death or serious morbidity
- 30 day postoperative death
- 30 day postoperative morbidity (NSQIP composite)
- 30-Day Postoperative Surgical Site Infection (SSI)

- 30-Day Postoperative Respiratory Complications (composite of pneumonia, intraoperative or postoperative unplanned intubation, pulmonary embolism, ventilation ≥48 hours)
- 30-Day Postoperative Pneumonia
- 30-Day Postoperative Urinary Tract Infection
- 30-Day Postoperative Renal Failure
- 30-Day Postoperative Venous Thromboembolism (deep vein thrombosis and/or pulmonary embolism)
- 30-Day Postoperative Stroke
- 30-Day Postoperative Myocardial Infarction
- 30-Day Intra- and Postoperative Transfusion
- 30-Day Postoperative Sepsis/Septic Shock
- 30-Day Postoperative Unplanned Return-to-Operating Room
- 30-Day Postoperative Unplanned Readmission
- Non-Home Discharge
- Failure to Rescue
- 30-Day Postoperative Cardiac Complications (NSQIP composite

Details on measurement of all secondary outcomes are in §E2.

SECTION D

STUDY DESIGN

SECTION OVERVIEW

D.1. OVERVIEW

The SECOND Trial is a two-arm cluster-randomized, pragmatic trial to test the effectiveness of a multi-component intervention to reduce burnout among residents training in accredited general surgery residency programs in the United States.

General surgery residency programs randomized to Control and Intervention arms will receive program-specific feedback reports on 3 basic resident outcomes: burnout, thoughts of attrition, and thoughts of suicide, as well as access to interventions related to suicidality.

General surgery residency programs randomized to Intervention will also receive a multi-component intervention consisting of (1) program-specific feedback on the learning environment and resident well-being, (2) a Wellness Toolkit containing ready-to-implement interventions for improving the learning environment and resident well-being, and (3) implementation support for adapting these interventions to the local context.

D.1.1. Justification of Cluster Randomization

The SECOND Trial is a cluster-randomized trial of general surgery residency programs and their residents to one of two study arms. The unit of randomization is the general surgery residency program. Individual residents will be assigned as a "cluster" to Control or Intervention based on the (random) study assignment of their residency program. This is because program policies and organizational features/structures to improve the learning environment and resident well-being are implemented at the organizational level. Once implemented, such measures are generally non-excludable – i.e. it cannot selectively benefit one resident but not another. Moreover, because some have argued and demonstrated that wellness (particularly suicidality) may be subject to social contagion,[25, 26] it is neither practically feasible nor statistically tenable to randomize individual residents within programs as any change in wellness may spill over and affect those in the same program regardless of study arm assignment under conditions of individual randomization.

D.1.2. Justification of Pragmatic Nature

The SECOND Trial is a pragmatic trial because programs will be offered feedback and strategies for improving the learning environment and resident well-being, but there will be no mandate by the Investigators to force Intervention programs to consider and/or act upon their feedback or prevent Control programs from acting upon theirs. Programs may attempt to improve their program measures at will, regardless of study arm. The only thing that distinguishes the Intervention from the Control Arm is the receipt of granular information on the learning environment, access to the Wellness Toolkit, and implementation support in the Intervention Arm.

D.2. INSTITUTIONAL REVIEW BOARD ASSURANCES

The ABSITE Survey protocol was reviewed by the Northwestern University IRB (NU IRB) as part of the FIRST Trial protocol and determined to be exempt from further review (IRB #STU00208523).

The Program Wellness Initiatives Inventory Survey protocol was reviewed by the NU IRB and approval was expedited (IRB #STU00209204).

The Program Tours component of the study protocol was reviewed by the NU IRB and approval was expedited (IRB #STU00208859).

The SECOND Trial protocol was reviewed by the NU IRB and deemed to be non-human subjects research and thus exempt from further review (IRB #STU00209709).

See **Appendix 2** for facsimiles of IRB documentation pertaining to the SECOND Trial.

Residency programs may elect to submit IRB applications to their local IRBs. Given that the trial itself was deemed to be non-human subjects research, NU IRB will not act as the IRB of record for external sites.

D.3. STUDY POPULATION

The study population for the SECOND Trial includes all general surgery residency programs in the United States that were accredited by the Accreditation Council for Graduate Medical Education (ACGME) as of April 2019.

D.4. STUDY ELIGIBILITY CRITERIA

D.4.1. Study Inclusion Criteria

Residency programs were eligible to enroll/participate in the SECOND Trial if they met the following study inclusion criteria:

- The residency program was in General Surgery.
- The residency program was ACGME-approved as of April 2019.

D.4.2. Study Exclusion Criteria

Residency programs were ineligible for participation in the SECOND Trial if they met the following exclusion criteria:

- The residency program was not in General Surgery.

D.5. RECRUITMENT AND ENROLLMENT

This section provides an overview of the procedures used to recruit eligible general surgery residency programs (see §D.4) into the SECOND Trial.

D.5.1. Sampling Frame

A list of all ACGME-approved General Surgery residency programs was provided by the ACGME. This list served as our "sampling frame."

D.5.2. Initial Contact

An initial contact email was sent to the program directors of all residency programs in the sampling frame. The initial contact email (1) provided a descriptive overview of the SECOND Trial, including key objectives of the study and answers to frequently asked questions, (2) invited them to attend one of two informational webinars and/or an in-person Town Hall meeting at the annual Association of Program Directors in Surgery (APDS) meeting. In both the webinars and the Town Hall meeting, additional details about the SECOND Trial were shared and opportunities to ask questions and/or give feedback were given. The initial contact email was signed by the co-Principal Investigators (YYH, KYB) and leadership from the ACS and APDS.

D.5.3. Informational Outreach

Two one-hour long informational webinars were conducted over the internet on April 4, 2019 and April 10, 2019. One one-hour long Town Hall meeting was conducted during the annual Association of Program Directors in Surgery meeting on April 2019. The informational webinars and the Town Hall meeting were led by the Principal Investigators. Attendees included one or more of the following representatives from the residency programs in our target sample: residency program directors and coordinators; leaders of surgical departments; educational and/or wellness leaders in surgical residency programs. The following content was covered:

- Study motivation
- Study objectives
- General design and endpoints
- Study eligibility
- Recruitment/enrollment
- Randomization
- Study arm descriptions – i.e., what participation entails
- Data collection
- Study timeline
- IRB exemption/approval
- Institutional endorsement/support – i.e., ACGME, ACS, ABS, APDS, Society of Surgical Chairs (SSC), the Association for American Medical Colleges (AAMC), and the Residents and Associates Society of the ACS (RAS-ACS)
- Question and answer
- Next steps

D.5.4. Enrollment Forms

On June 26, 2019, an enrollment email was sent to department chairs, program directors, program coordinators, and designated institutional officials at each program in our sampling frame. This email summarized the information that was presented during the informational webinars and Town Hall meeting, requested basic information about the residency programs, enumerated terms of participation, and requested signatures from the department chairs, program directors, and designated institutional officials. Programs were given until August 30 to remit completed enrollment forms. The deadline was then extended to October 31, 2019, given the considerable interest from programs who needed more time.

Terms of participation for all eligible residency programs were as follows:

- Agree to randomization to one of the two study arms
- Agree that if randomized to the intervention arm, it is entirely up to the program leadership to decide which, if any, interventions to implement
- Agree that their Program Director will participate in an initial brief phone call with the study team to learn about their program, if requested
- Agree that their Program Director will participate in a brief annual survey in order to establish which wellness interventions are underway or have been initiated at their program. (As was done for the FIRST Trial)
- As the reports provided by the SECOND Trial are for internal evaluation to improve wellness, agree not to advertise data any of the data provided by the SECOND Trial regarding resident wellness (e.g., burnout, attrition) in any public forum, including, but not limited to, social media, program website, resident recruitment/interview materials, and resident recruitment day activities
- Agree to not share access or content of Wellness Toolkit with other residency programs during the course of the Trial. All participating programs will receive the Toolkit at the end of the study
- Acknowledge that their residents will complete a brief survey annually for the SECOND Trial (As was done for the FIRST Trial)
- Agree that the American Board of Surgery (ABS) can share de-identified data with the SECOND Trial Coordinating Center, including (1) American Board of Surgery In-Training Examination (ABSITE) and (2) ABS written Qualifying Examination and oral Certifying Examination scores and pass rates (As was done for the FIRST Trial)
- Agree the Accreditation Council of Graduate Medical Education (ACGME) can share electronic data for years 2014-2024 with the SECOND Trial Coordinating Center, including (1) Annual Resident Survey data aggregated at program level (no individual identifiers); (2) Annual Faculty Survey data aggregated at program level (no individual identifiers); (3) Case Log de-identified, individual-level data (i.e., to examine resident case volumes for the surgical defined categories) with program identifiers; (4) any general descriptive information on programs (e.g., size, geography); (5) de-identified individual-level data on residents who did not graduate from their programs (e.g., demographics, year of departure, PGY level, and reasons for leaving if available) with program identifiers; (6) contact information for Program Director(s), Program Coordinator(s), Designated Institutional Official, and Chair for each program; and (8) other relevant program-level data that may be beneficial for the SECOND Trial. Programs will be identified via ACGME program identifier and no ACGME data will be transmitted if there are less than four responses per program per survey in any given year or the programs failed to meet the minimum compliance rate
- Agree to allow the American College of Surgeons National Surgical Quality Improvement Program (ACS NSQIP) to share de-identified data with the SECOND Trial Coordinating Center (As was done for the FIRST Trial)

- Agree to allow the Association of American Medical Colleges (AAMC) to share de-identified data with the SECOND Trial Coordinating Center, including (1) medical school graduate survey; (2) resident, faculty, leadership characteristics/demographics; and (3) other relevant program-level data that may be beneficial for the SECOND Trial

In 2020, an addendum to the agreement between ACS NSQIP and the hospital will be signed by each willing participating hospital. This addendum will allow transfer of ACS NSQIP data to the study team at Northwestern.

D.5.5. Follow-Up Contact

We maintain a log of all programs in our sampling frame that returned enrollment forms. On August 5, 2019, a reminder email was sent to program directors and department chairs for programs in the sampling frame that had not yet returned their forms. On August 13, 2019, a reminder email was sent to designated institutional officials for programs in the sampling frame that had not yet returned their forms. Starting on August 28, 2019, the co-PIs began contacting the remaining programs by email and/or text message to encourage them to enroll, ask about the barriers to enrollment, and offer facilitation (e.g., by extending the deadline, providing IRB or protocol materials). Based on feedback from New York State programs, on September 4, 2019, a clarifying email was sent to all unenrolled New York State programs to notify them that they were eligible for the SECOND Trial; unlike the FIRST Trial, New York State legislation does preclude participation. On September 4, 2019, the members of the Resident Advisory Committee were provided with a list of unenrolled programs so that they could contact residents that they personally know and encourage them to speak with their program directors about enrolling. On September 5, 2019, programs that had completed a Program Wellness Initiatives Inventory Survey, but not yet submitted an enrollment form were contacted. On September 9, 2019, the members of the Faculty Advisory Committee were provided with a list of unenrolled programs to similarly encourage enrollment. On September 25, 2019, a final reminder email was sent to unenrolled programs.

D.5.6. Enrollment Closure

Enrollment for the SECOND Trial officially closed on August 30, 2019, but was extended until October 31, 2019, given a large number of interested programs that were still in the process of obtaining internal approvals.

D.5.7. Enrollment Materials

Please see **Appendix 3** for facsimiles of recruitment materials (including, but not limited to contact emails, informational outreach materials, and the enrollment form/participation agreements).

D.6. STUDY UNITS

D.6.1. Unit of Randomization

The unit of randomization is the general surgery residency program. Eligible residency programs will be randomized to Control or Intervention.

D.6.2. Level of Observation

The unit of observation is the unit or entity upon which data were collected. The units of observation in this study are:

- Residency program
- Patient (secondary analysis)
- Resident (secondary analysis)

D.7. RANDOMIZATION PROCEDURES

We plan to use stratified cluster-randomization to allocate general surgery residency programs to the two study arms to help ensure balance of programs with respect to prior levels of resident burnout. We anticipate that there is a range of extant wellness activities across general surgery residency programs prior to the start of the SECOND Trial. The concern is that chance allocation may result in an imbalance of residency programs across study arms with respect to previous experience/activity in addressing resident wellness issues.

D.7.1. Cluster Unit of Randomization

The unit of randomization will be the general surgery residency program.

D.7.2. Definition of Randomization Strata

Using data from the 2019 ABSITE Survey, we will compute mean resident burnout scores (§C.2) within each residency program. Programs will then be stratified into quartiles based on mean resident burnout scores in 2019. Programs that do not have 2019 ABSITE data will be grouped into a fifth stratum ("no previous data").

The choice to define randomization strata on the basis of prior levels of resident burnout in programs was motivated by a desire to balance programs across study arms according to previous exposure to wellness activities and programming. We assume that extant levels of program engagement in issues of resident wellness will be reflected – at least to some degree – in mean levels of burnout within programs.

The number of strata was chosen based on a consideration of: avoiding thin strata; avoiding the inclusion of an excessive number of additional parameters in the final analytic models; and the ability to reduce between-program variance. Quartiles were chosen over tertiles because inclusion of quartiles in preliminary models showed that quartiles were slightly more effective in reducing between-program variance by (86% vs. 75%) compared to empty models (§D.8).

D.7.3. Randomization Procedure and Treatment Assignment

General surgery residency programs that enroll in the SECOND Trial will be grouped into four strata:

- Stratum 1 – Quartile 1, 2019 ABSITE program mean resident burnout scores
- Stratum 2 – Quartile 2, 2019 ABSITE program mean resident burnout scores
- Stratum 3 – Quartile 3, 2019 ABSITE program mean resident burnout scores
- Stratum 4 – Quartile 4, 2019 ABSITE program mean resident burnout scores
- Stratum 5 – *No data* on 2019 ABSITE program mean resident burnout score

Residency programs within each stratum will be assigned a unique, randomly-generated integer using a random-number generator. Residency programs within each stratum will be ordered in ascending order according to their randomly-assigned number. Four separate, blinded lists will be constructed containing the randomly-generated numbers corresponding to each program within each of the four strata. These blinded lists will be given to the Project Coordinator, who, with another Study Team member to be appointed, will alternately assign programs to Control or to Intervention. They will decide to start with control or intervention on the basis of a coin toss. An observer unaffiliated with the study will be present.

Once study arm assignments have been made for each of the four blinded lists, the blinded lists will then be sent back to the Statistical Analyst who will then unblind the lists. Final, unblinded residency program study arm assignments will be sent back to the Project Coordinator, so that participating programs may be notified of their study arm assignment. Study arm assignments will not be made public, but programs participating in the trial will be listed on the SECOND Trial website and www.clinicaltrials.gov.

D.8. SAMPLE SIZE/POWER CALCULATIONS

D.8.1. Data for Sample Size/Power Calculations

Data for power calculations came from the 2019 ABSITE Survey and are based on data from residents who were clinically active. Calculations were made for the primary endpoint only, burnout (§C.1 and §E.1).

In the 2019 ABSITE survey data, complete data on burnout survey items Q15A-Q15F and burnout scores were available for 5,962 clinically active residents from 300 programs. Clinically inactive residents were excluded, as were residents that had incomplete data on burnout survey items Q15A-Q15F. Program sizes, measured in terms of the number of respondents, ranged from 3 to 54, with a mean of 19.873 (SD=9.647).

Because we intend to randomize programs within strata defined on the basis of program-level mean burnout using prior data, we also used data from the 2018 ABSITE survey in order to define program-level strata for our preliminary models from which parameters were obtained for power calculations.

D.8.2. Outcome Measurement

As discussed in §E.1.2.3., our outcome measure in the SECOND Trial is an item response theory (IRT) derived measure of the latent burnout trait (theta) that has been linearly transformed into T-scores. This will be our outcome measure in our models, and this is the measure that we use for our power calculations.

A complete description of how T-scores were derived in the 2019 ABSITE data is given in §E. In 2019, the mean burnout T-score was 50.008, with a standard deviation of 9.500. Burnout T-scores ranged from a minimum of 32.078 to a maximum of 80.367.

D.8.3. Preliminary Models

Using ABSITE data from 2018, we computed IRT-derived T-scores for residents (2018 ABSITE survey items Q11A-Q11F). We then computed program-level mean burnout T-scores, and then stratified programs into quartiles based on these program-level mean burnout scores. These 2018 burnout quartiles were merged back to the 2019 ABSITE data. Programs that were missing in the 2018 ABSITE were grouped into a fifth stratum.

We estimated (a) a two-level ("empty") hierarchical linear model with random program intercepts; (b) an "unadjusted" hierarchical model adjusting only for program strata (quartile of 2018 burnout scores); and (c) an "adjusted" model that controlled for resident gender (female (reference male)); post-graduate year (PGY); resident relationship status; resident race (nonwhite (reference white)) and program quartile of burnout based on program-level mean burnout scores from 2018, program geographic region (Southeast, Midwest, Southwest, West (reference Northeast)); program type (community, military (reference academic)), and program size.

The intracluster correlation from the empty model was 0.082, suggesting that roughly 8% of total variance in resident burnout scores in 2019 were attributable to the program level. Controlling for program quartile of 2018 mean burnout scores in the "unadjusted" model reduced the program-level (Level-2) variance component by 75%, from 7.413 to 1.821, and reduced the ICC to 0.021. Including additional resident and program covariates reduced the Level-2 variance component by 86% compared to the empty model. The ICC in the adjusted model was 0.012 – i.e., approximately 1% of the variance in resident burnout scores was attributable to the program level after controlling for a program's prior year mean burnout T-score and resident/program characteristics.

The population-averaged mean burnout score estimated from the empty model was 49.753 (SD=9.137, square root Level 1 residual variance). The mean resident burnout score from the "unadjusted model" with controls for 2018 program mean burnout was 46.970 (SD=9.150). The mean resident burnout score from the adjusted model was 45.933 (SD=9.027, Level 1 residual).

Using the constant estimated from the adjusted model as an estimate of the mean burnout score in Control (45.933) and the standard deviation of the Level-1 error component as the estimate of the standard deviation (9.027) around mean burnout scores in Control and assuming $\text{alpha}_{\text{final}}$=0.0443 for the final analysis, 80% power and ICC=0.012, we computed minimal detectable differences for a variety of hypothetical sample sizes, as reported in EXHIBIT D.8-B.

The scenario yielding the smallest detectable difference in adjusted mean burnout scores would be one in which there are 300 programs total, 150 programs in each arm with a mean of 25 residents per program. In this scenario, given our assumptions, we expect to be powered at 80% to detect a relative difference in mean burnout scores of -1.46%.

In a conservative scenario with 100 programs randomized to each study arm (200 programs total) and an average of 15 residents per program, we expect to be powered at 80% to detect a relative difference in mean burnout scores of -2.22%.

In a moderate scenario with 110 programs randomized to each study arm (220 programs total) and an average of 20 residents per program, we expect to be powered at 80% to detect a relative difference in mean burnout scores of -1.87%.

EXHIBIT D.8.A. Preliminary Burnout Model for Power Calculations

		Coefficient (95% CI)	
	Empty Model	Unadjusted Model	Adjusted Model
Program Strata (ref: Stratum 1)			
Stratum 2	---	2.510 (1.643-3.377)	2.134 (1.261-3.007)
Stratum 3	---	3.442 (2.575-4.309)	3.076 (2.202-3.951)
Stratum 4	---	6.258 (5.384-7.131)	5.598 (4.719-6.477)
Stratum 5	---	0.273 (-0.885-1.430)	---
Female (ref: male)	---	---	2.064 (1.544-2.583)
PGY level (ref: PGY1)			
PGY2	---	---	0.479 (-0.276-1.233)
PGY3	---	---	-0.101 (-0.874-0.673)
PGY4	---	---	-0.607 (-1.393-0.179)
PGY5	---	---	-0.548 (-1.350-0.253)
Relationship status (ref: not in relationship)			
Married	---	---	0.265 (-0.408-0.938)
In a relationship	---	---	0.965 (0.279-1.651)
Divorced/widowed	---	---	-0.318 (-2.203-1.566)
Non-white (ref: white)	---	---	-1.512 (-2.056- -0.968)
Geographic region (ref: Northeast)			
Southeast	---	---	0.124 (-0.690-0.939)
Midwest	---	---	0.307 (-0.484-1.098)
Southwest	---	---	0.033 (-0.958-1.023)
West	---	---	0.064 (-0.299-1.582)
Program type			
Community	---	---	-0.069 (-0.810-0.672)
Military	---	---	-0.322 (-1.985-1.341)
Program size	---	---	0.125 (-0.056-0.306)
Constant	49.753 (49.356-50.150)	46.970 (46.325-47.615)	45.933 (44.461-47.405)
Variance components			
Level-2	7.413 (5.627-9.767)	1.821 (1.010-3.285)	1.010 (0.417-2.446)
Level-1	83.483 (80.459-86.622)	83.721 (80.686-86.870)	81.485 (78.303-84.795)
Intracluster correlation	0.082	0.021	0.012
N residents	5962	5962	5075
N programs	300	300	237
Likelihood ratio test vs. linear model (x2)	172.20	18.29	6.94
P value	P<0.001	P<0.001	P=0.004

Data: 2019 (version 2) and 2018 ABSITE Survey data. The dependent variable is an IRT-derived burnout T-score based on 6 survey items corresponding to emotional exhaustion and depersonalization subscales (2019 ABSITE survey items Q15A-Q15F, see §E.1 for details). Program strata refers to program quartile based on program mean 2018 ABSITE burnout T-scores. The 5th stratum contains programs that had missing 2018 ABSITE data and were thus not ranked in a quartile. However, this stratum dropped out of the adjusted model due to collinearity.

EXHIBIT D.8.A. Preliminary Burnout Model for Power Calculations

		Coefficient (95% CI)	
	Empty	Unadjusted Model	Adjusted Model

	Model		
Program Strata (ref: Stratum 1)			
Stratum 2	---	2.510 (1.643-3.377)	2.134 (1.261-3.007)
Stratum 3	---	3.442 (2.575-4.309)	3.076 (2.202-3.951)
Stratum 4	---	6.258 (5.384-7.131)	5.598 (4.719-6.477)
Stratum 5	---	0.273 (-0.885-1.430)	---
Female (ref: male)	---	---	2.064 (1.544-2.583)
PGY level (ref: PGY1)			
PGY2	---	---	0.479 (-0.276-1.233)
PGY3	---	---	-0.101 (-0.874-0.673)
PGY4	---	---	-0.607 (-1.393-0.179)
PGY5	---	---	-0.548 (-1.350-0.253)
Relationship status (ref: not in relationship)			
Married	---	---	0.265 (-0.408-0.938)
In a relationship	---	---	0.965 (0.279-1.651)
Divorced/widowed	---	---	-0.318 (-2.203-1.566)
Non-white (ref: white)	---	---	-1.512 (-2.056- -0.968)
Geographic region (ref: Northeast)			
Southeast	---	---	0.124 (-0.690-0.939)
Midwest	---	---	0.307 (-0.484-1.098)
Southwest	---	---	0.033 (-0.958-1.023)
West	---	---	0.064 (-0.299-1.582)
Program type			
Community	---	---	-0.069 (-0.810-0.672)
Military	---	---	-0.322 (-1.985-1.341)
Program size	---	---	0.125 (-0.056-0.306)
Constant	49.753 (49.356-50.150)	46.970 (46.325-47.615)	45.933 (44.461-47.405)
Variance components			
Level-2	7.413 (5.627-9.767)	1.821 (1.010-3.285)	1.010 (0.417-2.446)
Level-1	83.483 (80.459-86.622)	83.721 (80.686-86.870)	81.485 (78.303-84.795)
Intracluster correlation	0.082	0.021	0.012
N residents	5962	5962	5075
N programs	300	300	237
Likelihood ratio test vs. linear model (x2)	172.20	18.29	6.94
P value	P<0.001	P<0.001	P=0.004

Data: 2019 (version 2) and 2018 ABSITE Survey data. The dependent variable is an IRT-derived burnout T-score based on 6 survey items corresponding to emotional exhaustion and depersonalization subscales (2019 ABSITE survey items Q15A-Q15F, see §E.1 for details). Program strata refers to program quartile based on program mean 2018 ABSITE burnout T-scores. The 5th stratum contains programs that had missing 2018 ABSITE data and were thus not ranked in a quartile. However, this stratum dropped out of the adjusted model due to collinearity.

EXHIBIT D.8.B Minimal Detectable Differences for Primary Analysis under Varying Sample Size Scenarios

Number of Programs (Clusters) per Arm	Number of Residents per	Usual Care Mean Burnout	Estimated Intervention	Minimal Detectable Difference	Relative Difference (%)

	Program (Mean Cluster Size)		Mean Burnout		
150	25	45.93	45.26	-0.67	-1.46%
150	20	45.93	45.19	-0.74	-1.61%
150	15	45.93	45.10	-0.83	-1.81%
125	25	45.93	45.19	-0.74	-1.61%
125	20	45.93	45.12	-0.81	-1.76%
125	15	45.93	45.02	-0.91	-1.98%
110	25	45.93	45.14	-0.79	-1.72%
110	20	45.93	45.07	-0.86	-1.87%
110	15	45.93	44.96	-0.97	-2.11%
100	25	45.93	45.10	-0.83	-1.81%
100	20	45.93	45.03	-0.90	-1.96%
100	15	45.93	44.91	-1.02	-2.22%
NOTES: Minimal detectable difference from adjusted estimate of mean burnout scores, assuming mean burnout T-scores in Usual Care=45.933, SD in both arms=9.027, ICC=0.012, alpha$_{final}$=0.0443 and power=80%, based on preliminary estimates in EXHIBIT D.8-B. Relative difference is expressed as a percent of Usual Care mean.					

D.9. STUDY SAMPLE

A total of 209 residency programs enrolled in the SECOND Trial.

Exhibit D.9 shows the distribution of SECOND Trial residency program participants as of November 8, 2019.

Exhibit D.9. SECOND Trial Sample as of November 8, 2019

UNIT	Control N	Intervention N	Total N
General Surgery Residency Programs	106	103	209
Mean Burnout T-score[a] (SD)	50.48 (3.07)	49.98 (3.88)	-
Program Type			
Academic	55	52	107
Community	46	48	94
Military	3	2	5
Mean Number of Residents (SD)	34.08 (16.77)	33.88 (15.95)	-

NOTE: [a] Based on 2019 ABSITE data.

D.10. STUDY ARM DESCRIPTIONS

D.10.1. Intervention

General surgery programs randomized to Intervention will receive:

- Program-specific feedback information on the learning environment as well as resident wellness
- A multi-component Wellness Toolkit of structured interventions for programs to implement to improve their learning environment and resident wellness.

A distinguishing feature of the intervention is that the study team intends to leave it to programs to decide whether and how to act upon the information the Investigators provide. It is the intent of the Investigators to provide information to programs to facilitate appropriate action given a program's specific needs or areas for improvement.

The intervention is expected to be delivered in winter 2019 – 2020.

D.10.2. Control

General surgery residency programs randomized to Control will be expected to conduct "business as usual." At the behest of Program Directors, Control programs will receive program-specific feedback on resident wellness outcomes (burnout, suicidality, and thoughts of quitting), as well as suicide-related interventions, to ensure that no program will be at a disadvantage with respect to preventing resident suicide.

D.11. DATA COLLECTION PROCEDURES

D.11.1. ABS Data

By enrolling in the FIRST Trial, residency programs agreed to grant the Study Team access to de-identified data from the SECOND Trial Resident Survey (**Appendix 4**). The SECOND Study Team designed a brief, closed-ended survey administered annually to all surgical residents taking the January ABSITE; however, the administration of the survey as part of the examination is under the sole purview of the ABS. Data collection and data processing is also executed by the ABS.

The ABS will give the Study Team access to a de-identified dataset containing coded, private information, ABSITE scores, and responses to items in the SECOND Trial Resident Survey Addendum. This de-identified dataset will *not* include any direct personal identifiers (i.e., personal names, addresses, or direct institutional identifiers). The dataset will include analytic identifiers for individual respondents and an analytic institutional identifier to permit appropriate statistical methods for analyzing the data to account for clustering of data. The ABS will also give the Study Team access to a dataset containing program-level ABS Qualifying Examination and Certifying Examination pass rates.

D.11.2. Program Wellness Initiatives Inventory Survey

On April 3, 2019, personalized links to the Program Wellness Initiatives Survey were emailed to all ACGME-accredited general surgery program directors. A reminder was sent to non-responders on June 12, 2019. A reminder to enrollees who had not yet completed the survey was sent following enrollment closing and prior to release of the program-specific report. Completion of the survey is a prerequisite for SECOND Trial enrollment.

The Program Wellness Initiatives Inventory Survey queries program directors about existing activities, resources, education, training, policies, and perceptions of problems (**Appendix 4**). It will be distributed via Qualtrics annually to assess what programs, both Control and Intervention, are doing to address well-being.

D.11.4. ACGME Data

ACGME Case Log data and ACGME Annual Resident and Faculty Survey Data will be obtained from the ACGME. The dataset will include an analytic institutional identifier to permit appropriate statistical methods for analyzing the data to account for clustering of data.

D. 11.5. ACS NSQIP Data

The hospital of primarily affiliated with the enrolled residency programs will be asked in 2020 to grant the Study Team access to their hospitals' ACS NSQIP data.

A detailed, official description of the ACS NSQIP data can be found here: http://site.acsnsqip.org/participant-use-data-file/

D.11.6. OTHER DATA

Hospital-level data from the AHA *Annual Survey of Hospitals* and from CMS *Annual Payment Update Files (PUF)* may be merged at the hospital level to ACS NSQIP data to obtain additional information on hospital characteristics (e.g., bed size, nurse staffing ratios, resident-to-bed ratios, overall case mix index).

D.12. STUDY ADHERENCE AND DATA SAFETY MONITORING

D.12.1. Study Adherence

This is a pragmatic trial. Monitoring to ensure strict adherence to Study Arm conditions will not be undertaken because doing so would threaten the external validity of this study.

However, we will attempt to assess program adherence using the annual Program Wellness Initiatives Inventory Survey (**Appendix 4**) and semi-structured interviews with program directors for the following purposes:

- To undertake "per-protocol" or "as-treated" analyses to estimate the local average treatment effect of increased wellness activities/resources on patient and resident outcomes (to complement our main intent-to-treat analyses)

- To ascertain the demand for personalized data and/or ready-to-implement interventions – i.e., do program directors take advantage of these resources to make targeted changes in their programs?

Adherence may be based on metrics that include but are not limited to:

- Usage of the Wellness Toolkit website
- Completion of the annual Program Wellness Initiatives Inventory Survey
- Implementation of wellness interventions
- Utilization of wellness-intervention implementation support
- Sharing of report data with residents and faculty
- Stability of program leadership
- Buy-in and commitment of program leadership toward improvement of resident well-being and learning environment
- Allocation of resources to intervention implementation

D.12.2. Data Safety Monitoring

D.12.2.1. Data Safety Monitoring Board (DSMB)

We convened a Data Safety Monitoring Board (DSMB) to independently review study progress at regular intervals. The purpose for establishing a DSMB is to serve as an independent review body to evaluate the results of a planned Interim Analysis that is scheduled for January 2021. The Interim Analysis, described in greater detail in §F.8, will use SECOND Trial Resident Survey data collected from participating programs from 2019 to 2021 to compare resident burnout in programs randomized to Intervention vs. those in programs randomized to Control. The DSMB was charged with determining whether early stopping rules for the SECOND Trial had been met, and if so, to terminate the trial out of consideration for resident wellness.

D.12.2.2. DSMB Composition

The SECOND Trial DSMB was designed to have 3 individuals as follows:

- One (1) academic surgeon with a surgical education focus
- One (1) academic surgeon with a health services research focus
- One (1) statistician or biostatistician

D.12.2.3. Selection of DSMB Members

D.12.2.3.1. Eligibility Criteria

- Members of the SECOND Trial Study Team are ineligible to serve on the DSMB
- Members of the SECOND Trial Advisory Committees are ineligible to serve on the DSMB

D.12.2.3.2. Selection Procedure

Potential DSMB candidates will be vetted by the SECOND Trial Oversight Committee. The Principal Investigators will select 3 candidates to serve as DSMB members.

D.12.2.4. DSMB Responsibilities and Functions

The SECOND Trial DSMB was charged with the following responsibilities and functions:

- Review the SECOND Trial Protocol/Statistical Analysis Plan
- Review the SECOND Trial Interim Analysis Report
- Attend an Interim Conference Call
- Pending results in the SECOND Trial Interim Conference Call and Report, recommend to the SECOND Trial Study Team/Principal Investigator one of the following actions
    - Discontinue the study
    - Continue the study according to protocol

D.12.2.5. DSMB Meetings

D.12.2.5.1. Scheduled Meetings

The DSMB will convene by web conferencing twice during the study.

The Interim Meeting will be held in May 2021, following the Interim Analyses.

The Final Meeting will be held in May 2023 once the Final Analyses are completed.

D.12.2.5.2. Quorum

Given the small size (3 members) of the DSMB, all members must be present for the meetings/calls. Voting and discussions prior to voting will be held “behind closed doors” – i.e., in the absence of the Principal Investigators, Study Team, and/or Oversight Committee.

D.12.2.5.3. Voting

DSMB Members must vote on recommendations to the SECOND Trial Study Team/Oversight Committee. Outcomes will be based on a simple majority rule. No abstentions from voting will be permitted. Study Team and Oversight Committee members will not be permitted to attend the voting session or meeting.

D.12.2.5.4. Procedures for Making Recommendations

DSMB recommendations must be submitted in writing to the Principal Investigators within seven days of the voting outcome. The Principal Investigators will then apprise other Study Team members and the Oversight Committee of the DSMB's recommendation.

D.12.2.5.5. Meeting Format

Each of the two scheduled meetings will have an *Open Session* and a *Closed Session.*

- Open Session. Each DSMB Meeting will begin with an Open Session during which the Principal Investigators, Study Team Members, and possibly Oversight Committee members will attend and participate. During the Open Session, the Principal Investigator and other members of the Study Team and/or Oversight Committee may apprise DSMB members of current study status/progress updates/evolving issues. An appointed Study Team member will take minutes of the Open Session.

- Closed Session. Each DSMB Meeting will end with a Closed Session during which DSMB members will confer amongst themselves. The Interim Analysis vote (DSM Charter §C.1.) will be held during the Closed Session.

D.12.2.5.6. Confidentiality

DSMB members will treat all study materials and communications as confidential unless otherwise specified.

D.12.2.5.7. Conflicts of Interest

DSMB members may not be members of the ACS, ACGME or ABS. DSMB members may not be faculty affiliates of Northwestern University.

SECTION E

MEASURES

SECTION OVERVIEW

E.1. MEASUREMENT OF PRIMARY ENDPOINT

The primary outcome of the SECOND Trial will be resident burnout.

E.1.1. Primary Endpoint Data

The primary outcome will be resident burnout, measured on the basis of six ABSITE survey items adapted from the Maslach Burnout Inventory (MBI).[19] The items measure:

- [Q15A] *feeling morning fatigue*
- [Q15B] *feeling strained by working with people*
- [Q15C] *feeling emotionally drained*
- [Q15D] *treating patients impersonally*
- [Q15E] *becoming more callous*
- [Q15F] *not caring about patients' outcomes*

In the 2019 survey, each item was scored on a 7-point Likert-type scale with the following categories:

1. Never
2. A few times a year
3. Once a month or less
4. A few times a month
5. Once a week
6. A few times a week
7. Every day

E.1.2. Item Response Theory (IRT) Scoring of Burnout (Primary Endpoint)

Although the MBI was intended to provide a near-continuous measure of burnout, scoring of MBI and the measurement of burnout has been inconsistent in the published literature.[17] The SECOND Trial will implement scoring based on an item-response-theory (IRT) graded-response model analysis of the 2019 ABSITE burnout data.

E.1.2.1. Item-Response Theory (IRT) Graded Response Model (GRM)

Item-response theory (IRT) assumes that burnout is an unobservable individual-level latent trait. Resident responses to survey items designed to assess levels of burnout reflect varying amounts of this latent trait. IRT models the relationship between how residents respond to these survey items (i.e. the probability of a particular response) and their underlying level of "burnout." For example, IRT would posit that residents with low levels of the latent burnout trait will have a higher probability of responding "never" to Item Q15C ("I feel emotionally drained from my work") whereas residents with higher levels of burnout will have a low probability of responding "never", and a higher probability of responding "a few times a week" to Item Q15C.

Samejima's graded response model (GRM) is an IRT model that is widely used for ordered, polytomous variables. In our burnout application, each of the 6 MBI burnout survey items had 7 response categories. Thus, the IRT GRM fitted a set of six (7-1=6), 2-parameter IRT models for each survey item. The GRM estimates 6 category response functions which characterize how levels of the underlying latent burnout trait are related to response probabilities for each of the 7-1 response categories. The graph of these category response functions is called the boundary characteristic curve. The GRM constrains the slope of all boundary characteristic curves to be constant across response categories within an item, but slopes can vary across items. The slopes of the boundary characteristic curves correspond to discrimination (the steeper the slope, the greater the discrimination between respondents with more or less of the underlying latent construct). The GRM estimates separate location parameters for each response category. These parameters relate the amount of the underlying latent construct results in a 50% chance of selecting a particular response category or higher. The location parameters correspond to the "difficulty" of a particular item.

IRT models assume unidimensionality, local independence, and monotonicity.[27] Unidimensionality of the data implies that the six MBI items all measure a single construct and that inter-item covariances are driven by a single latent burnout factor and random error. Local independence implies that, conditional on the latent factor, the six burnout survey items are otherwise independent – i.e. there is no residual correlation among the items. Monotonicity implies that the relationship between levels of the latent construct and the probability of a keyed response is unidirectional

If these assumptions hold in the data,[27-30] IRT models allow characterization of each survey item according to its level of discrimination (capacity to differentiate between respondents with varying levels of the latent trait) and difficulty. Individual-level burnout scores can also be obtained from the IRT models that reflect levels of the latent burnout trait. These latent scores, typically referred to as "theta" (θ) have a mean of 0 and a standard deviation of 1.

E.1.2.2. Rationale for using IRT-Derived Burnout Measure

We chose to use IRT-derived measures of burnout as our measure of the primary outcome in the SECOND Trial for several reasons.

- IRT provides a seamless framework for us to allow the psychometric performance of each MBI item in our population and to characterize the discrimination and difficulty of each item, as well as generate a theoretically-grounded, empirically-based composite score of burnout that does ***not*** rely on arbitrary assumptions or choice of thresholds and cutoffs for dichotomizing measures. Raw scores computed as a sum of responses assume that each item is of equal discrimination and difficulty.[31]
- IRT-derived measures have desirable scaling properties. Theta is a near-interval measure and can be interpreted similarly to z-scores and may be preferable to arbitrarily dichotomizing items. When used as an outcome variable, theta is more likely to be (approximately) normal than an additive raw score which is at the ordinal level of measurement.
- Although we do not pursue it formally in the SECOND Trial, IRT paves the way for more in-depth psychometric evaluation of burnout in terms of differential item function (DIF); that is, whether survey items perform differently in different subgroups of the population. Future research could investigate whether certain demographic or academic subgroups of the resident population respond to MBI items differently, which would result in spurious subgroup differences in levels of burnout.
- When IRT assumptions hold, IRT parameter estimates are item-invariant and sample-invariant. Although the topic of IRT linking and scale reconciliation is a complex methodological field, IRT provides the greatest potential for generating comparable measures of burnout across samples, across variants of survey instruments, and over time. Thus, we adopt IRT-based measurement of

burnout in this Trial in part to address the broader issues of incommensurability in burnout assessment across studies.

E.1.2.2. IRT GRM Analysis of Burnout Using 2019 ABSITE Data

We used data from the 2019 ABSITE Survey to perform an IRT GRM analysis of the 6 burnout items in the 2019 ABSITE data in order to develop our outcome measure.

In this work, 2019 ABSITE residents who did not report being in a clinically active year (Item Q1==1) were excluded. This work was performed on complete cases. Therefore, residents with any missing data on items Q15A-Q15F were also excluded. (Preliminary work retained residents were clinically inactive, as well as residents with some missing data. Results were similar.)

E.1.2.2.A. Evaluation of IRT Assumptions

We first assessed whether IRT assumptions were met in the 2019 ABSITE data (survey items Q15A-Q15F).

Unidimensionality. Unidimensional IRT assumes that a multi-item scale reflects a single, unitary construct (synonyms that we may use interchangeably include “trait,” “latent factor,” and “factor”). We adopted three common approaches to assess unidimensionality of the burnout scale in the 2019 ABSITE data: (a) Cronbach’s alpha to measure internal consistency; (b) exploratory factor analysis; and (c) confirmatory bifactor analysis.

(a) *Cronbach’s alpha.* The internal consistency of the 6 burnout survey items was assessed using Cronbach’s alpha (α), and was found to be $\alpha = 0.900$. TABLE 1 shows the item-test correlations and item-rest correlations for each item. The final column estimates the value of alpha for a 5-item scale if a particular item were removed. Cronbach’s alpha ranges from 0 to 1. Rules of thumb suggest that values of 0.70-0.80 may be in the lowest acceptable range, and that values ≥0.95 are desirable. Thus, with an alpha of 0.900, the 6-item scale demonstrated satisfactory inter-item reliability. As evident in the final column, reliability could not be improved by the removal of any ill-fitting item as alpha became smaller with removal.

(b) *Exploratory factor analysis (EFA).* Unidimensionality is commonly assessed using exploratory factor analysis (EFA) and/or confirmatory factor analysis (CFA).[32, 33] Because all 6 survey items are polytomous, we performed an EFA on a polychoric correlation matrix. The eigenvalue of a factor is a quantitative measure of the amount of variance in the nine burnout items that a factor explains. Generally, one retains factors with eigenvalues greater than 1. TABLE 2 reports the eigenvalues estimated for each factor from the EFA, as well as the proportion of total variance and cumulative variance accounted for by the factors. In this analysis, a single factor had an eigenvalue greater than 1, and was thus retained. Factor 1 had an eigenvalue=4.101; in other words, this single factor explained roughly as much variance in the 6 variables as 4 variables might. This factor explains roughly 96% of the variance in these items.

TABLE 3 shows factor loadings for the six items. For illustrative purposes, two sets of loadings are shown. All items loaded strongly onto the first factor, and had very weak loadings on the second factor. Factor loadings are analogous to correlations. Thus, items Q15a-f were highly correlated with Factor 1, but generally uncorrelated with Factor 2.

We computed the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy, a test to determine the proportion of variance in the 6 items may be "explained" by latent factors. This is a diagnostic procedure to determine the suitability of the items for factor analysis. KMO values range from 0-1.00. Rules of thumb for interpreting KMO values are: unacceptable [0.00-0.49]; miserable [0.50-0.59]; mediocre [0.60-0.69]; middling [0.70-0.79]; meritorious [0.80-0.89]; and marvelous [0.90-1.00]. The overall KMO of the 6 items was 0.871, or "meritorious" – suggesting that factor analysis was appropriate for these data.

A second diagnostic test that was performed was Bartlett's test of sphericity, which tests the null hypothesis that the 6-item correlation matrix is an identity matrix. An identity matrix is one in which all off-diagonal cells are all zeros. This would suggest that items are unrelated to one another, and therefore unsuitable for factor analysis. The chi-square value for Bartlett's test was 23971 with 15 degrees of freedom (corresponding to the number of off-diagonal cells in the correlation matrix –i.e. number of inter-item correlations) and $p<0.0001$. This test also supported the use of EFA on these data.

(c) *Confirmatory bifactor analysis*. We also performed a confirmatory bifactor analysis of the 6 items to confirm that the factor structure of the 6-item burnout scale was "sufficiently unidimensional" to support IRT.[34, 35] The following three confirmatory factor structures were evaluated:

- Single factor model – all items were posited to load onto a single latent factor (burnout)
- Two correlated-factors model – emotional exhaustion and depersonalization are conceptualized as separate-yet-related constructs
- Bifactor model – all items were posited to load onto a single general factor as well as one of two subdomain factors (emotional exhaustion and depersonalization)

Schematic diagrams of these three factor structures are shown in FIGURE 1.

Because our underlying data consisted of ordered, polytomous variables, these models were estimated in MPlus (8.2) using weighted least mean squares adjusted estimation (WLSMV). To assess and compare these models, we examined traditional fit statistics and threshold values:

- Root mean square error of approximation (RMSEA) with values <0.08 for "adequate" fit and <0.05 for "good" fit[35, 36]
- Comparative fit index (CFI) with values >0.90 for acceptable/good fit[35, 36]
- Tucker-Lewis Index (TLI) with values >0.90[35]
- Standardized root mean residual (SRMR) with values ≤0.08 for acceptable/good fit[36]

Essential unidimensionality was assessed in the bifactor model by examining the relative strength of item loadings on the general factor vis a vis subdomain factors. Loadings ≥0.30 were considered to be "salient".[35] We considered factor loadings to be suggestive of essential unidimensionality if all item loadings on the general factor ≥0.30. Preferably, general factor loadings would be salient and larger than subdomain factor loadings. However, so long as general factor loadings are salient even if they are of equal or smaller magnitude than subdomain factor loadings, the general factor may still be considered to be essentially unidimensional despite the existence of "well-defined" subdomain factors.[35] In addition to inspecting loadings, we also examined the explained common variance (ECV) associated with the general factor, which reflects the fraction of total variance common to the 6 items that is accounted for by the general factor.[37] In the absence of consensus regarding a threshold value for ECV, we considered values ≥0.50 to suggest unidimensionality. Omega-H is another diagnostic measures for assessing unidimensionality. Omega-H measures the fraction of variance in the overall burnout score that can be attributable to the general factor.[38] Rodriguez, Reise and Haviland (2016) suggest a threshold value of Omega-H ≥0.80 to indicate essential unidimensionality of a construct.

FIGURE 2 shows the estimated loadings from the bifactor model. The magnitude of general factor loadings range from 0.652 (0.009) for Item q15a, to 0.951 (0.007) for Item q15e. These magnitudes exceed 0.30[35] and suggest the presence of a fairly well-defined general factor even if there may be a subfactor present. TABLE 4 provides additional support for the essential unidimensionality of the burnout construct. Focusing on the 6-item burnout scale, the fit statistics that we had chosen as the basis for model evaluation indicated that the bifactor model appeared best suited to the data. This model, compared to the single factor and two correlated factors model, had the lowest RMSEA and SRMR values and were both were below the threshold of 0.08. The bifactor model also had the highest CFI and TLI values, which both exceeded 0.90.

Taken together, Cronbach's alpha, exploratory, and confirmatory factor analyses suggest support for treating the 6-item burnout scale as essentially unidimensional.

Local Independence. Local independence is related to unidimensionality but evaluation of this assumption in polytomous IRT models is not yet well supported in software. Edelen and Reeve (2007) suggest inspecting the residual covariance matrix from a one-factor CFA to examine whether there is any evidence of "excess" covariance among items, but do not suggest any formal tests.[32, 33]

Monotonicity. To examine whether the relationship between response categories and the latent burnout score from the EFA exhibits monotonicity, FIGURES 1A-F plot the mean latent burnout score estimated from the EFA within each response category for each of the 6 burnout items.[39] We have also plotted burnoutθ estimated from the IRT model within each response category. In each of the 6 figures, we see that mean burnout scores increase across the ordered response categories with lower burnout scores associated with responses in the less frequent range, and higher burnout scores associated with responses in the higher frequency range.

We also performed a Mokken scale analysis of the 6 burnout items to assess overall scalability as well as checks on the assumption of monotonicity. All item-level Loevinger scaling coefficients (*H*) were above the 0.50 threshold, indicating "strong" scalability.[40, 41] With minvi values set at 0.30 (default) and alpha=0.05, the Mokken scaling procedure and Loveinger scaling procedure found no significant or non-significant deviations from monotonicity for any of the 6 burnout items. Trace plots of item category response rates plotted against item rest scores (total score – item score) also provide visual confirmation that there are no violations of the assumption of monotonicity in these data for the 6 burnout items (see Appendix to Section E1 for results and trace plots).

Summary. Together, these assessments suggest that the key assumptions for IRT are upheld in our data.

E.1.2.2.B. IRT GRM Analysis

IRT GRM Item Discrimination and Location Parameters. The IRT GRM discrimination and location parameters are presented in TABLE 4 for each of the 6 burnout scale items. Baker (2001, Table 2-4) suggests that discrimination parameters in the range of 1.35-1.69 indicate "high" discrimination and anything greater than 1.70 indicates "very high" discrimination. According to this rule-of-thumb, the 6 burnout survey items demonstrated "very high" discrimination.[42]

The location parameter of a response category curve estimate the value of burnoutθ that is associated with a 50% probability of selecting that response category or a higher category. Greater values indicate greater difficulty – that is, greater levels of the latent burnout trait. Thus, if we compare the top-box category in Q15A (*feeling morning fatigue*) to Q15F (*not caring about patients' outcomes*) we see that burnout$\theta_{7,Q15A}$=1.816 for but that burnout$\theta_{7,Q15F}$ =3.055. A 50% probability of responding "every day" to Q15A (*feeling morning fatigue*) requires lower levels of burnout compared to Q15F. A 50% probability of responding "every day" to Q15F (*not caring about patients' outcomes*) asesrequires higher levels of burnout. Therefore, Q15F might be said to be a more "difficult" item.

FIGURES 4A-4F are graphs the category response functions for each burnout survey item. The discrimination parameter corresponds to the slope of the response curves while the location parameter shifts each curve across the range of theta.

FIGURES 5A-5F show category characteristic curves for each item in the burnout scale. For each of the 7 response categories in a given survey item, the category characteristic curve plots the probability that a respondent will select that particular response category across the range of latent burnout trait. Points along each category characteristic curve at a given value of burnoutθ add up to 1.00.

Item and Test Information Functions. FIGURE 6 shows the IRT GRM item information functions for each of the 6 items in the burnout scale. Each line in this figure corresponds to a survey item, and each curve plots the amount of information the item provides across the range of theta values (latent burnout construct). The most informative items are Q15C, Q15D, and Q15E. The 6-item scale appears to provide greatest information on respondents with slightly higher-than-average levels of burnout. This is also seen in the overall test information function in FIGURE 7, which is the sum of individual information functions. Again, the 6-item burnout scale is most informative over above-average levels of burnout.

Test Characteristic Curve. As mentioned previously, IRT produces a non-linear, logistic re-scaling of the raw data and provides a way of translating latent burnout levels to expected raw scores. This is shown in FIGURE 8 which is the test characteristic curve for the 6-item burnout scale. Here, expected raw score is plotted against values of the latent burnout trait. Actual raw scores are overlaid against the latent burnout trait to allow visual inspection of overall fit. The density of observed scores are close to the test characteristic curve, suggesting generally adequate fit.

E.1.2.3. IRT Scoring of Burnout Using 2019 ABSITE Data

Theta is the IRT-derived measure of the latent burnout trait and has a mean of 0 and a standard deviation of 1.

Theta can, and is often linearly transformed into what is known as a "T-score" according to the formula:

$$T - Score = 50 + (10 \times \theta)$$

T-scores range from 0-100, and are often rounded to integers for reporting. T-scores are also interval measures and more approximately normally distributed. To illustrate the distributional advantages of raw summative scores, theta, and T-scores, histograms of each measure are shown in FIGURES 9A-9C. T-Scores are simply a linear transformation of theta, and thus their distributions are identical. Departures from normality are due to kurtosis, not skewness. The D'Agostino et al. test (*Stata* `sktest`) for skewness shows that while all three measures are kurtotic (T-score/theta p=0.010, raw burnout score p<0.001), T-scores/theta are not skewed (T-score/theta skewness=-0.003 (reference value 0), p=0.933) whereas raw burnout scores are right-skewed (skewness=0.698 (reference value 0), p<0.001). Non-skewness of IRT-derived measures is obviously desirable when it comes to parametric modeling.

For purposes of illustration only, FIGURE 10A shows a scatterplot of raw (summed) burnout scores against T-scores (a plot against theta would be identical).

E.1.2.3. IRT Scoring of Burnout in the SECOND Trial

Our primary outcome in the SECOND Trial is burnout. Our primary measure of burnout will be IRT-derived T-scores. Power calculations for this trial were executed on the basis of T-scores.

Evaluations and final academic reporting of the results of this trial will be made using T-scores.

We anticipate reports issued to programs as part of the Trial to be made also on the basis of T-scores.

TABLE 1. Internal Consistency of the 6-Item Burnout Scale

Burnout Item	Item-Test Correlation	Item-Rest Correlation	α (Item Removed)
Q15a *Feeling morning fatigue*	0.800	0.693	0.892
Q15b *Feeling strained by working with people*	0.825	0.744	0.883
Q15c *Feeling emotionally drained*	0.867	0.796	0.874
Q15d *Treating patients impersonally*	0.846	0.771	0.879
Q15e *Becoming more callous*	0.861	0.783	0.877
Q15f *Not caring about patients' outcomes*	0.718	0.623	0.900

Items response categories were: 1-Never; 2-A few times a year; 3-Once a month or less; 4-A few times a month; 5-Once a week; 6-A few times a week; 7-Every day. Cronbach's α for the 6 items was 0.902.

TABLE 2. Factor Eigenvalues from an Exploratory Factor Analysis of 6 Burnout Scale Items

Factor	Eigenvalue	Proportion Variance	Cumulative Variance
Factor 1	4.101	0.961	0.961
Factor 2	0.455	0.107	1.067
Factor 3	-0.020	-0.005	1.062
Factor 4	-0.065	-0.015	1.047
Factor 5	-0.097	-0.023	1.025
Factor 6	-0.105	-0.025	1.000

TABLE 3. Factor Loadings from an Exploratory Factor Analysis of 6 Burnout Scale Items

Item	Factor 1	Factor 2
Q15a *Feeling morning fatigue*	0.779	0.324
Q15b *Feeling strained by working with people*	0.817	0.179
Q15c *Feeling emotionally drained*	0.864	0.296
Q15d *Treating patients impersonally*	0.865	-0.286
Q15e *Becoming more callous*	0.862	-0.157
Q15f *Not caring about patient outcomes*	0.769	-0.353

TABLE 4. Factor Structure Evaluation of 6-Item Burnout Scale

Model Statistic	Single Factor	Correlated Factors	Bifactor Model
9-Item Burnout Scale			
Root mean square error of approximation (RMSEA)	0.469	0.112	0.094
Comparative fit index (CFI)	0.704	0.985	0.992
Tucker-Lewis Index (TLI)	0.606	0.978	0.984
Standardized root mean residual (SRMR)	0.175	0.037	0.025
Explained common variance (ECV)‡	---	---	0.522
omegaH‡	---	---	0.748
Percent uncontaminated correlations (PUC)	---	---	0.750
6-Item Burnout Scale			
Root mean square error of approximation (RMSEA)	0.257	0.126	0.066
Comparative fit index (CFI)	0.961	0.992	0.999
Tucker-Lewis Index (TLI)	0.935	0.985	0.996
Standardized root mean residual (SRMR)	0.051	0.020	0.006
Explained common variance (ECV)‡	---	---	0.671
omegaH‡ (General Factor)	---	---	0.861
Percent uncontaminated correlations (PUC)	---	---	0.600

N=6040, 2019 ABSITE survey data. Models were estimated using MPlus 8.3 using its weighted least square mean and variance adjusted estimator (WLSMV) for the analysis of ordered categorical variables. RMSEA<0.05 indicates “good fit” (<0.08 “adequate fit”). CFI>0.90, TLI>0.90 are generally accepted as indicative of adequate/good fit. SRMR<0.08 indicates good/adequate fit. [‡ Dueber, D. M. (2017). Bifactor Indices Calculator: A Microsoft Excel-based tool to calculate various indices relevant to bifactor CFA models. https://dx.doi.org/10.13023/edp.tool.01 [Available at http://sites.education.uky.edu/apslab/resources/]]

TABLE 4. Item Response Theory Graded Response Model Discrimination and Location Parameters for 6-Item Burnout Scale

Burnout Survey Item	**Discrimination**	**Location Parameters (Standard Error)**					
		≥2	**≥3**	**≥4**	**≥5**	**≥6**	**=7**
Q15a *Feeling morning fatigue*	2.335 (0.054)	-1.566 (0.033)	-0.565 (0.022)	-0.134 (0.020)	0.503 (0.021)	0.864 (0.023)	1.816 (0.037)
Q15b *Feeling strained by working with people*	2.778 (0.068)	-0.371 (0.020)	0.288 (0.019)	0.746 (0.021)	1.265 (0.027)	1.758 (0.034)	2.594 (0.059)
Q15c *Feeling emotionally drained*	3.153 (0.079)	-0.991 (0.024)	-0.144 (0.018)	0.294 (0.018)	0.875 (0.021)	1.253 (0.025)	2.042 (0.040)
Q15d *Treating patients impersonally*	3.161 (0.086)	-0.289 (0.019)	0.325 (0.018)	0.715 (0.020)	1.236 (0.025)	1.669 (0.032)	2.406 (0.051)
Q15e *Becoming more callous*	3.252 (0.082)	-0.600 (0.020)	0.059 (0.018)	0.435 (0.018)	0.924 (0.021)	1.284 (0.021)	1.884 (0.036)
Q15f *Not caring about patients' outcomes*	2.207 (0.064)	0.369 (0.021)	1.017 (0.027)	1.441 (0.033)	1.885 (0.042)	2.329 (0.054)	3.055 (0.084)

NOTES: N=5962. 2019 ABSITE Data, complete cases only (Q15A-Q15F). Clinically inactive residents excluded.

FIGURE 1A-1C. Schematic Diagrams of Factor Structure Models

Figure 1A. Single Factor Model

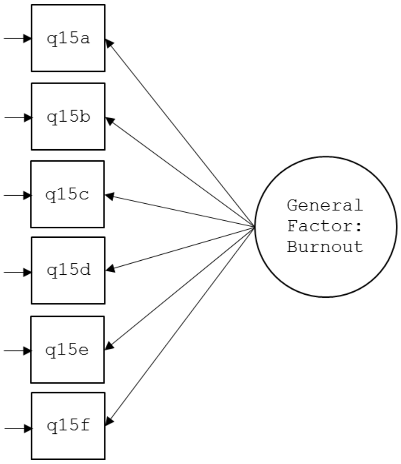


Figure 1B. Two Correlated Factors Model

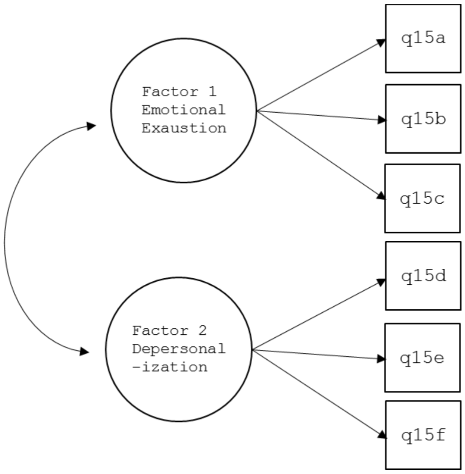


Figure 1C. Bifactor Model

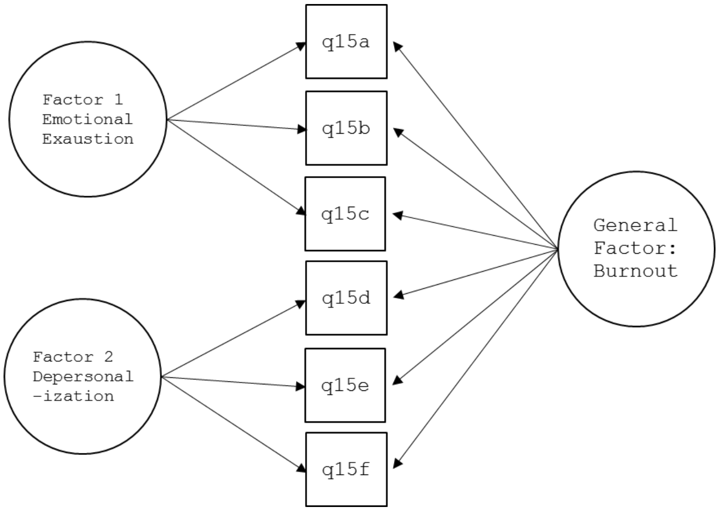
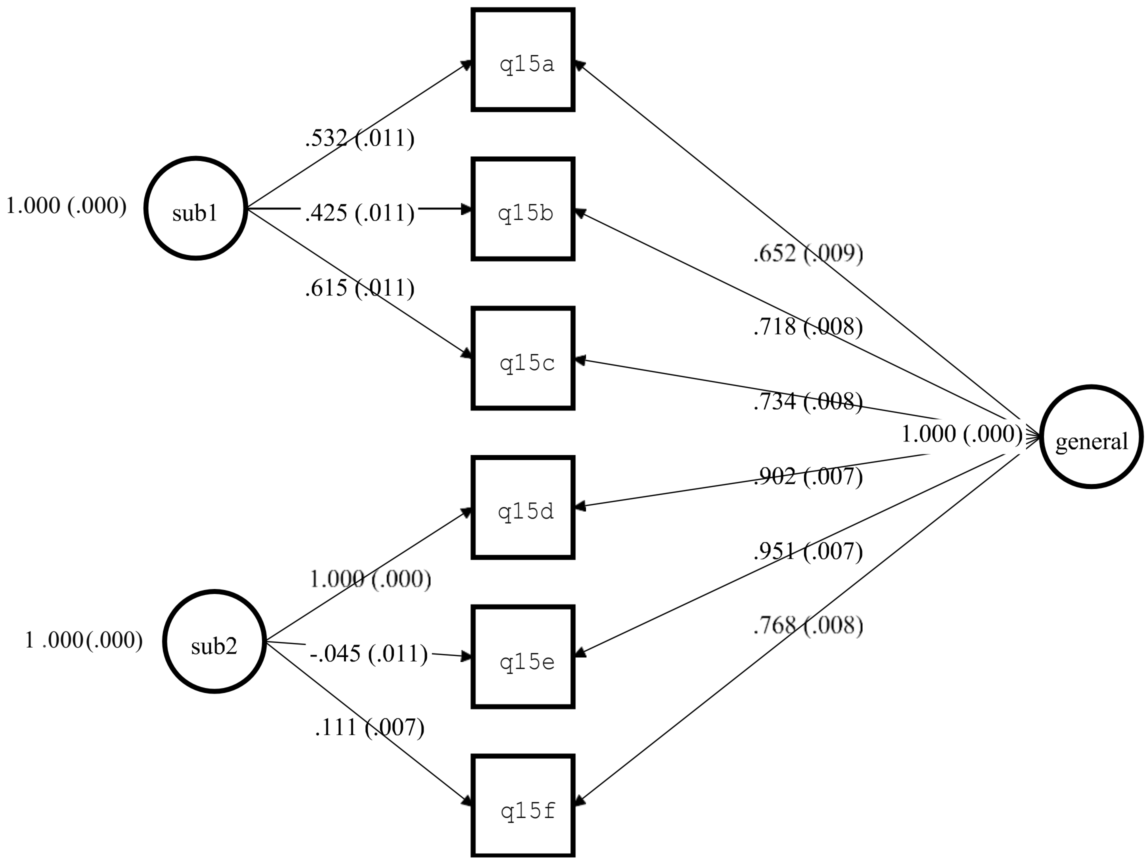


FIGURE 2. Confirmatory Bifactor Model and Loadings – Six-Item Burnout Scale



and

FIGURE 3A-F. Mean EFA/IRT-estimated Latent Burnout Trait Scores, by Burnout Response Categories

Commented [YH1]: Change all of these to paraphrases on page 30

Commented [DA2R1]: Done

Figure 3A. Item Q15A "Feeling morning fatigue"

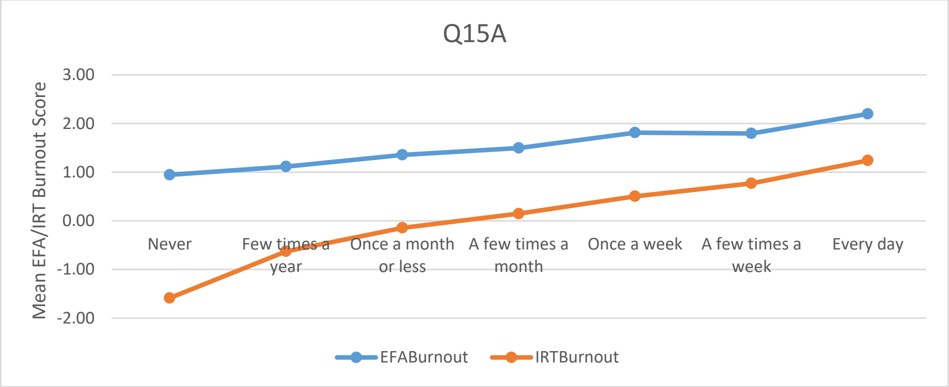


Figure 3B. Item Q15B "Feeling strained by working with people"

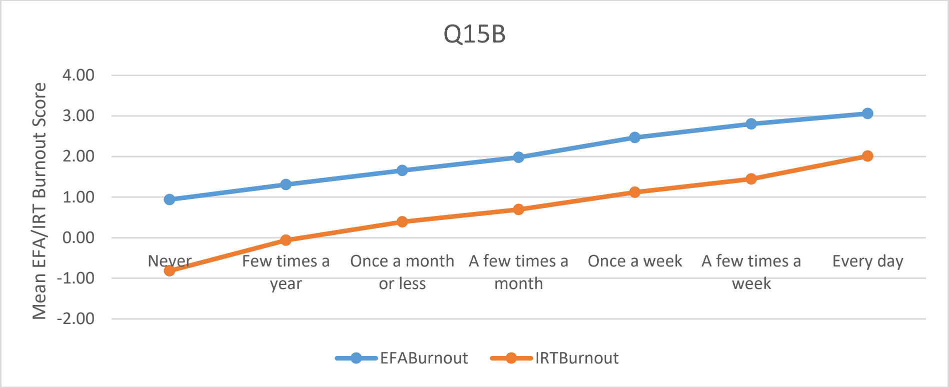


Figure 3C. Item Q15C "Feeling emotionally drained"

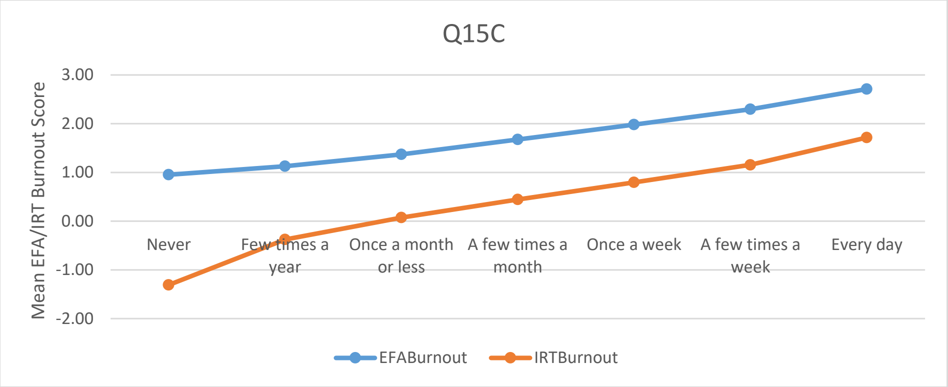


Figure 3D. Item Q15D "Treating patients impersonally"

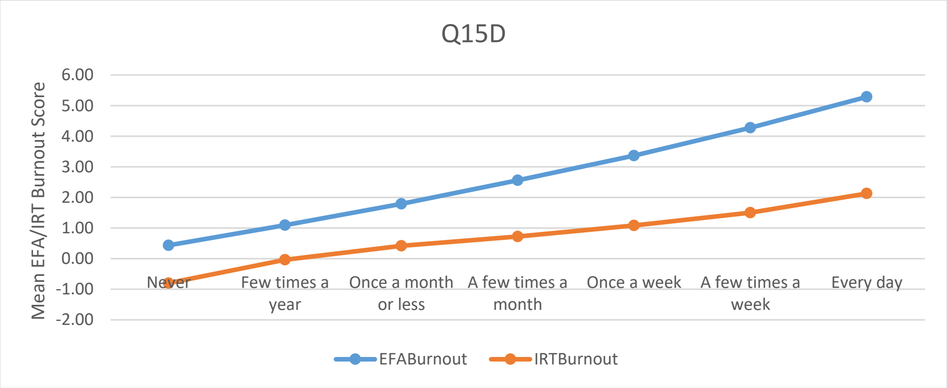


Figure 3E. Item Q15E "Becoming more callous"

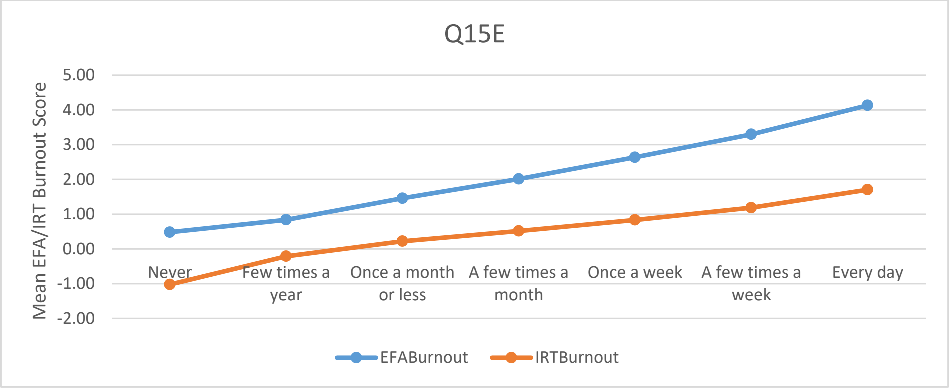


Figure 3F. Item Q15F "Not caring about patient outcomes"

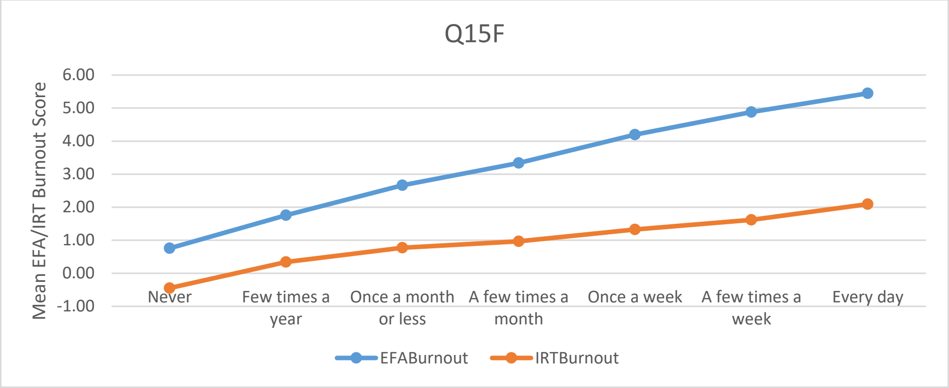
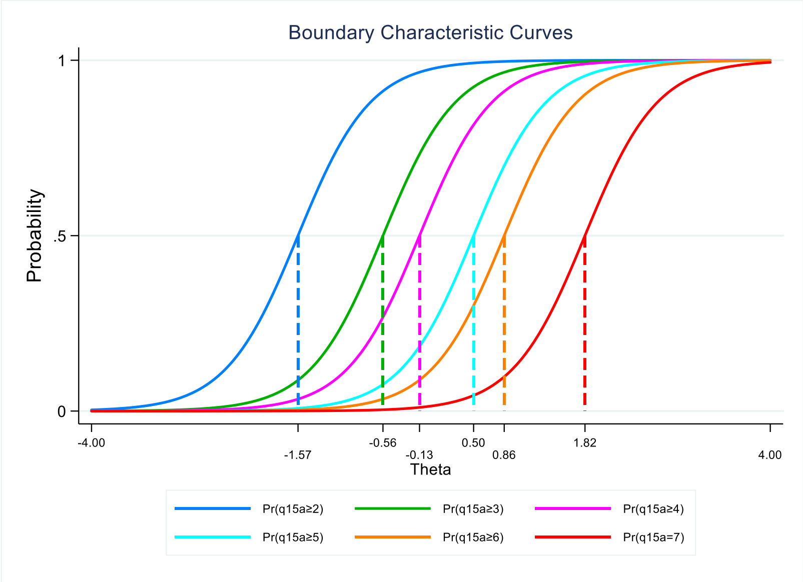


FIGURE 4A-F. IRT GRM Boundary Characteristic Curves

Figure 4A. Item Q15A "Feeling morning fatigue"



Commented [DA3]: •[Q15A] *feeling morning fatigue*
•[Q15B] *feeling strained by working with people*
•[Q15C] *feeling emotionally drained*
•[Q15D] *treating patients impersonally*
•[Q15E] *becoming more callous*
•[Q15F] *not caring about patients' outcomes*

Figure 4B. Item Q15B “Feeling strained by working with people”

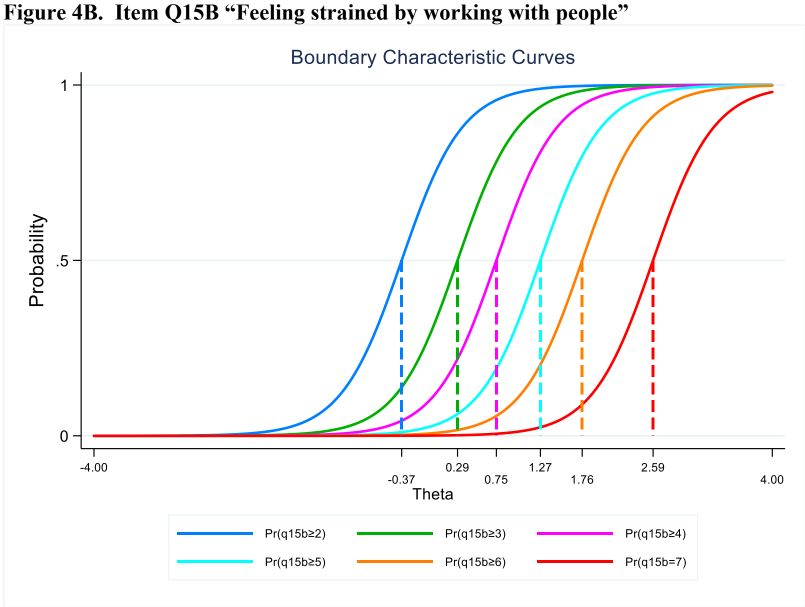


Figure 4C. Item Q15C “Feeling emotionally drained”

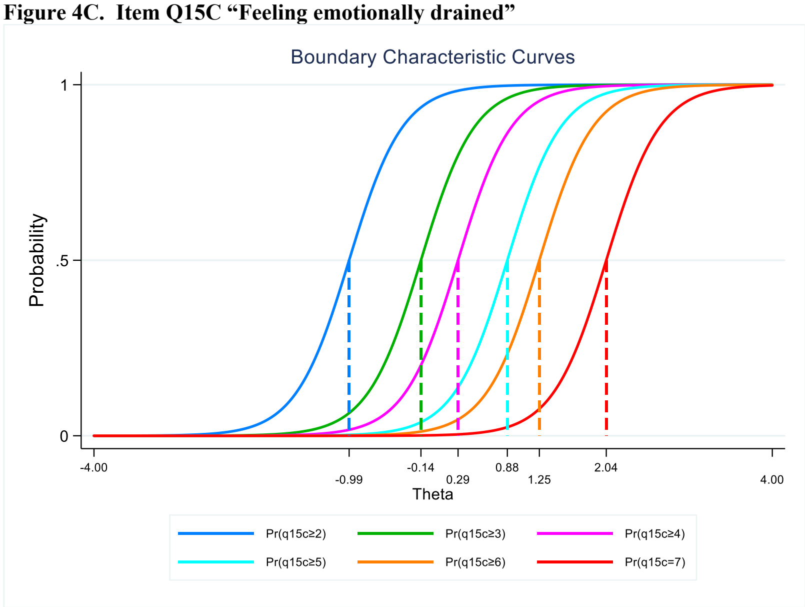


Figure 4D. Item Q15D “Treating patients impersonally”

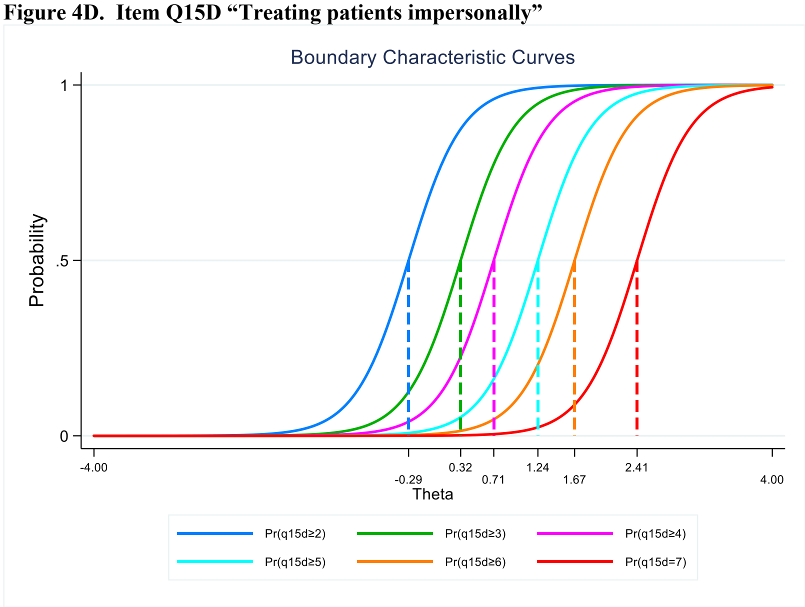


Figure 4E. Item Q15E “Becoming more callous”

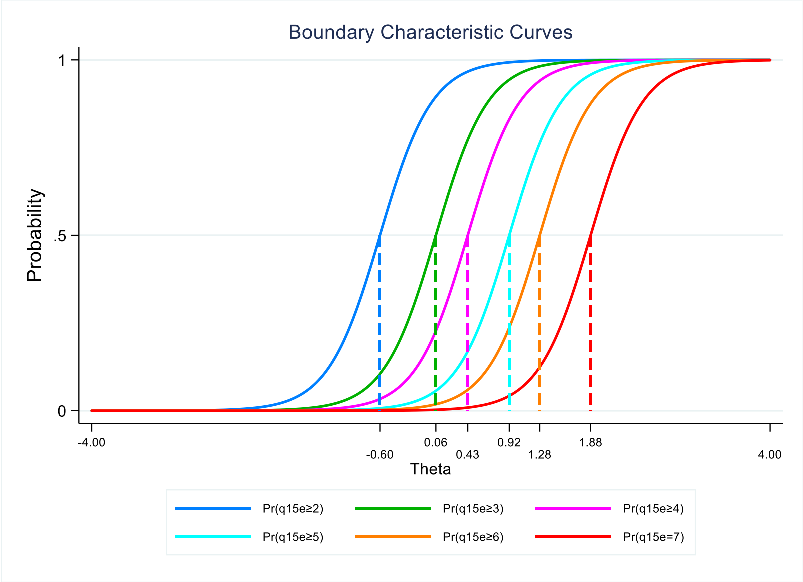


Figure 4F. Item Q15F "Not caring about patients' outcomes"

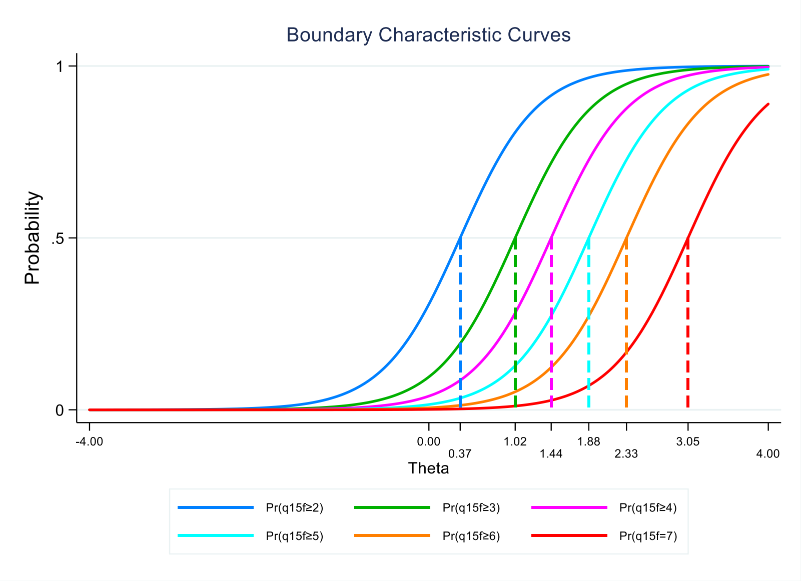
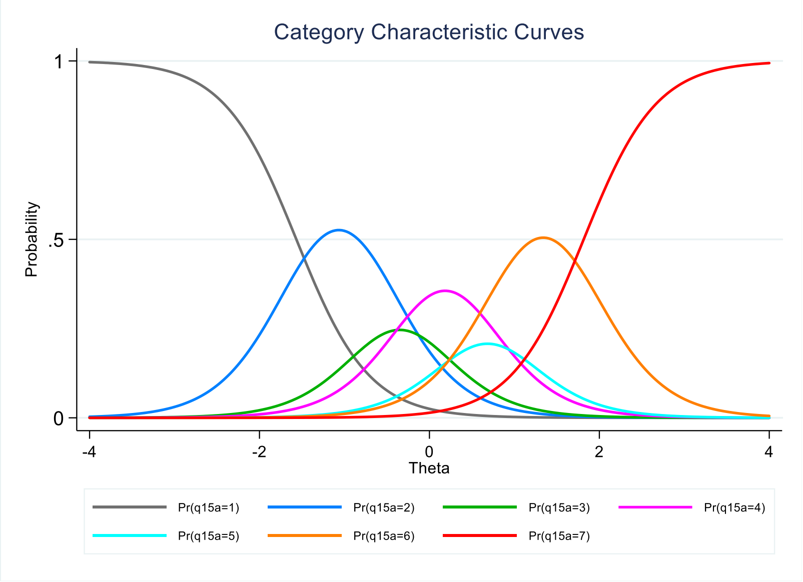


FIGURE 5A-F. IRT GRM Category Characteristic Curves

Figure 5A. Item Q15A "Feeling morning fatigue"



Commented [DA4]: •[Q15A] *feeling morning fatigue*
•[Q15B] *feeling strained by working with people*
•[Q15C] *feeling emotionally drained*
•[Q15D] *treating patients impersonally*
•[Q15E] *becoming more callous*
•[Q15F] *not caring about patients' outcomes*

Figure 5B. Item Q15B "Feeling strained by working with people"

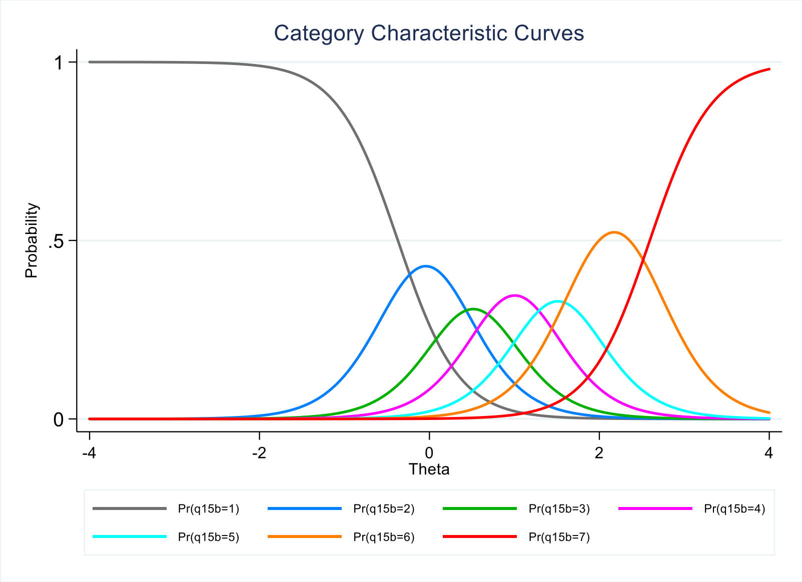


Figure 5C. Item Q15C "Feeling emotionally drained"

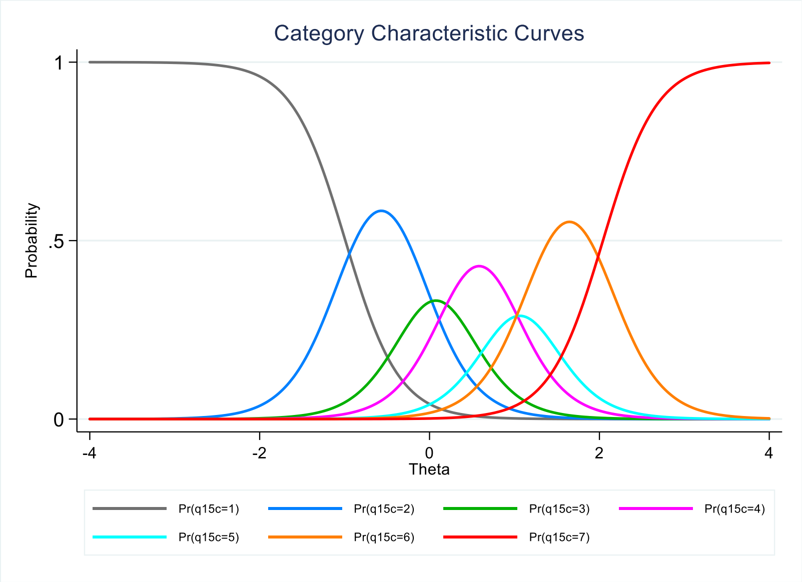


Figure 5D. Item Q15D "Treating people impersonally"

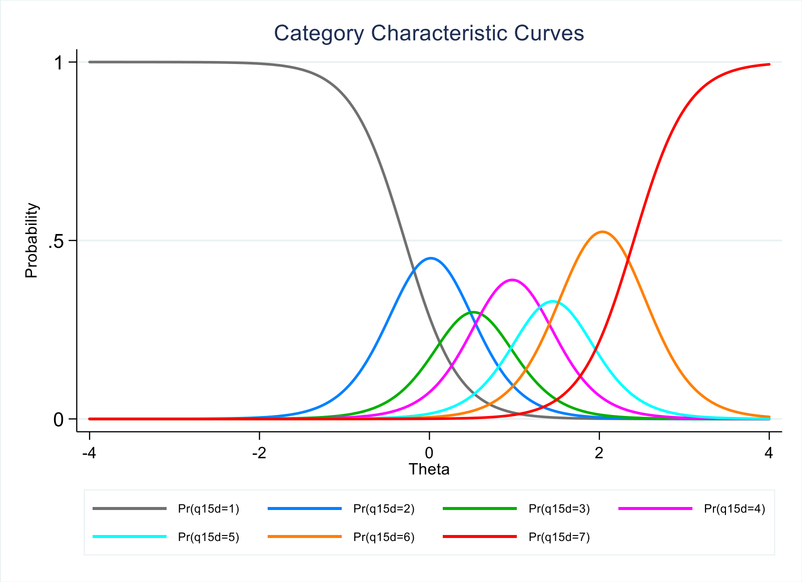


Figure 5E. Item Q15E "Becoming more callous"

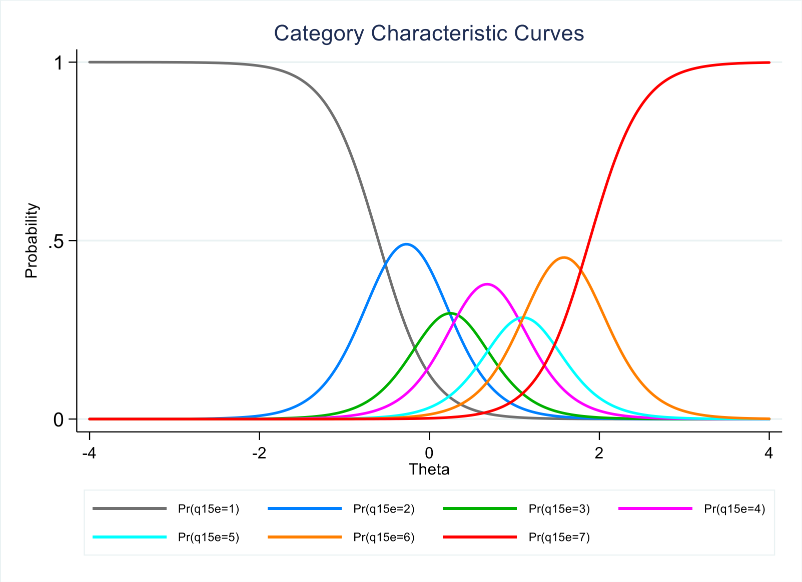


Figure 5F. Item Q15F "Not caring about patients' outcomes"

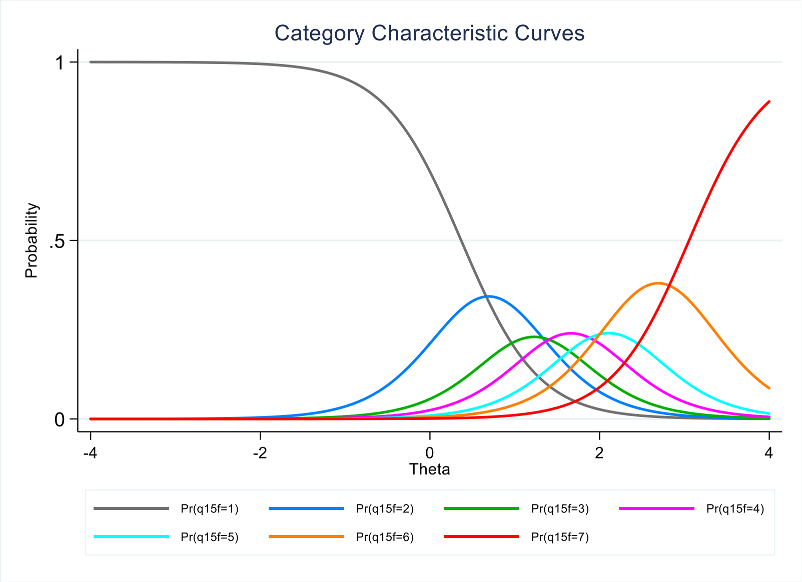


FIGURE 6. IRT GRM Item Information Functions

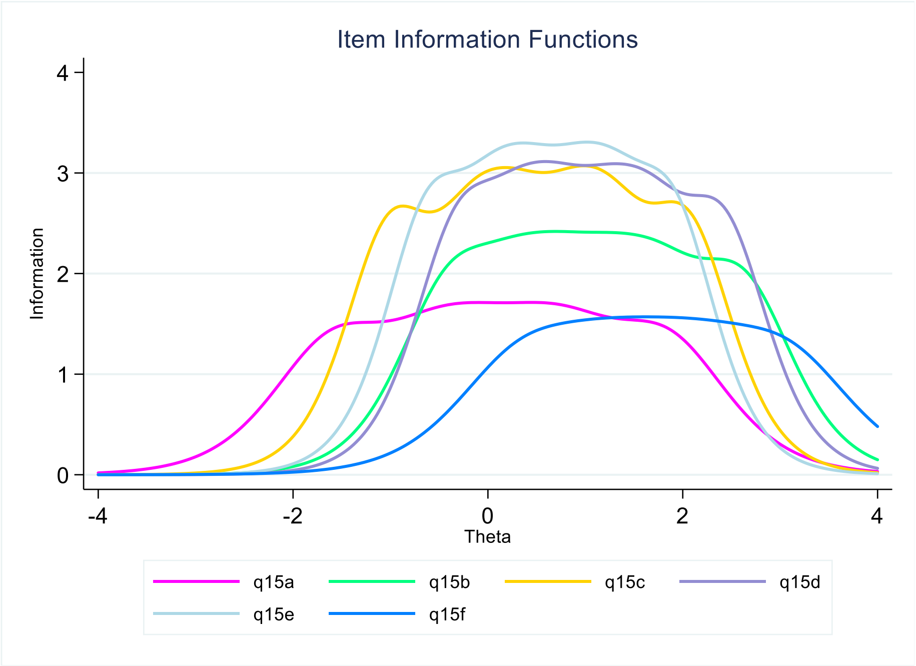


FIGURE 7. IRT GRM Test Information Function

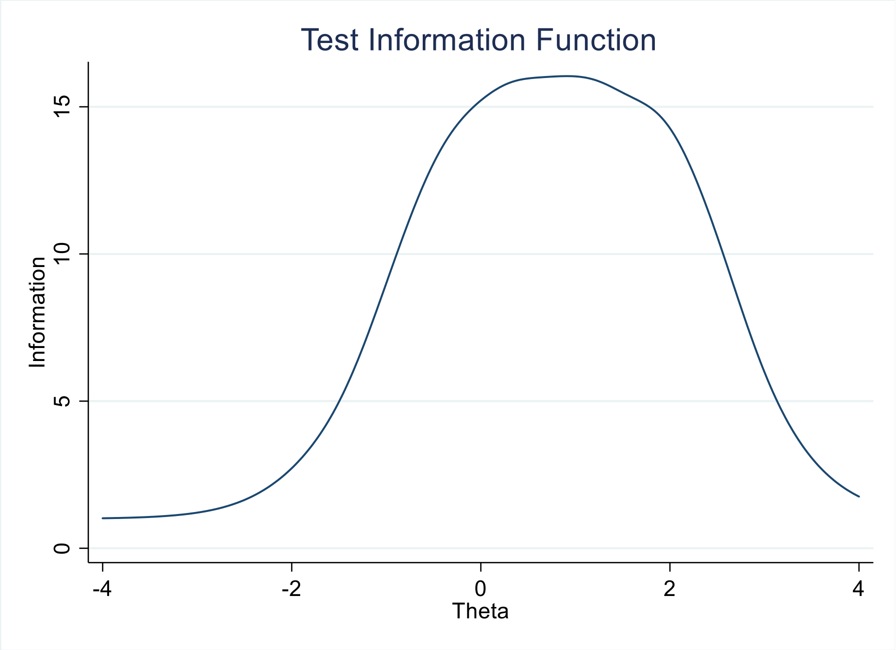
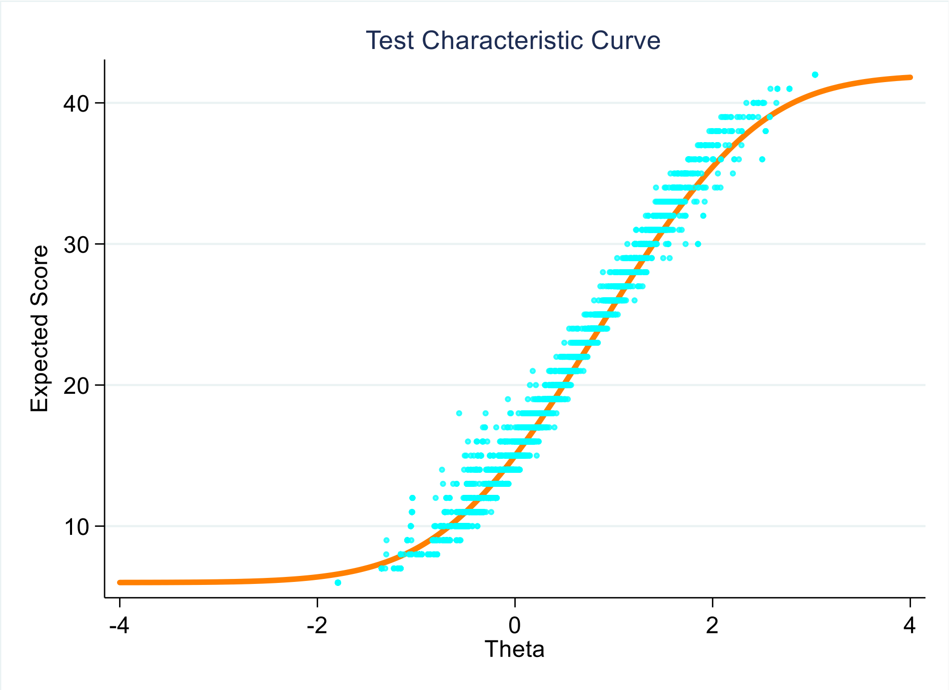


FIGURE 8. Test Characteristic Curve, with Raw Score Overlay



FIGURES 9A-C. Burnout Score Histograms

Figure 9A. Histogram of 2019 ABSITE Raw (Summed) Burnout Scores

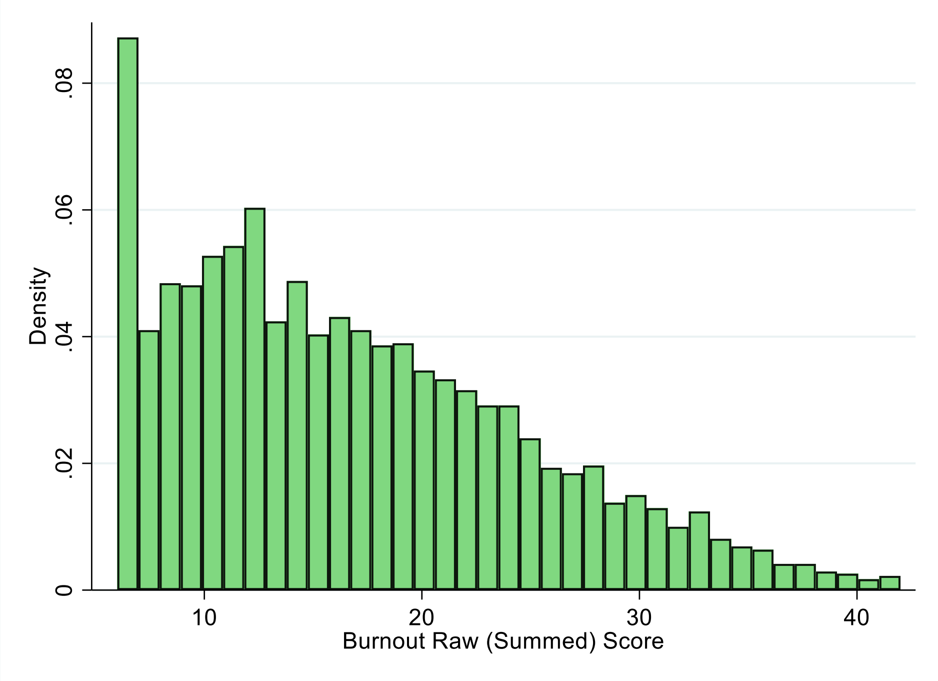


Figure 9B. Histogram of 2019 ABSITE Theta Burnout Scores (IRT GRM Latent Burnout Trait)

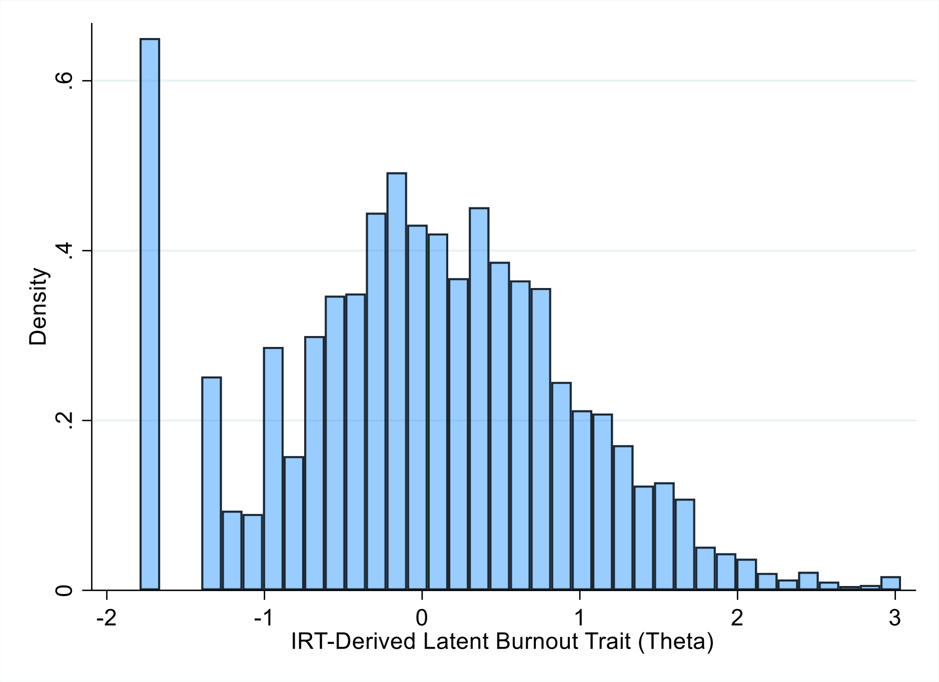
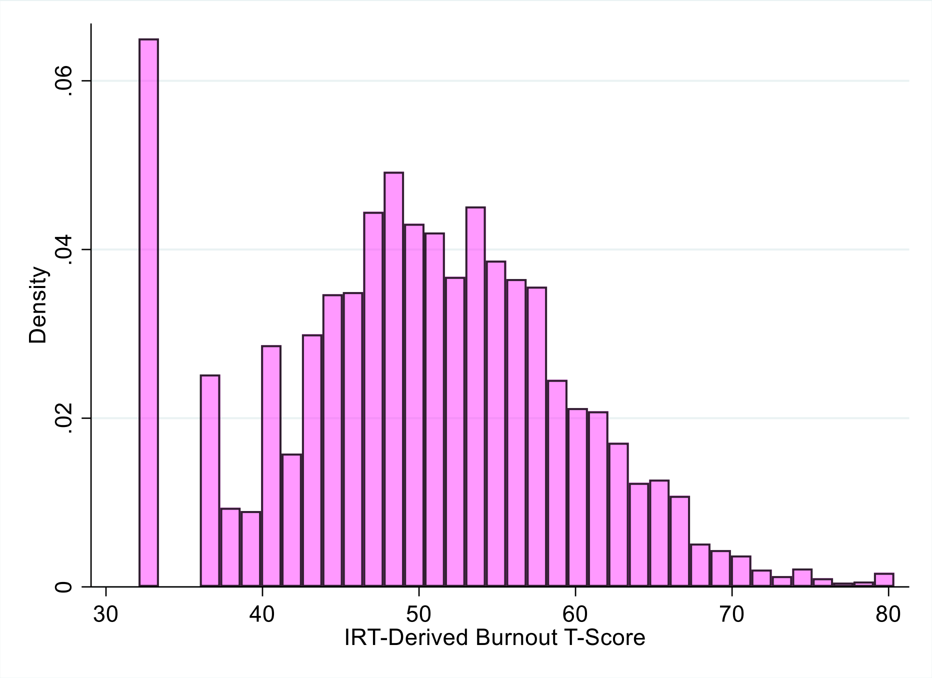


Figure 9C. Histogram of 2019 ABSITE Burnout T-Scores (IRT GRM Linear Transformation of Theta)



TABLES 10A-B. Scatter Plots of Burnout Measures

Table 10A. Scatterplot of Raw Burnout Scores Against IRT-Derived Burnout T-Scores

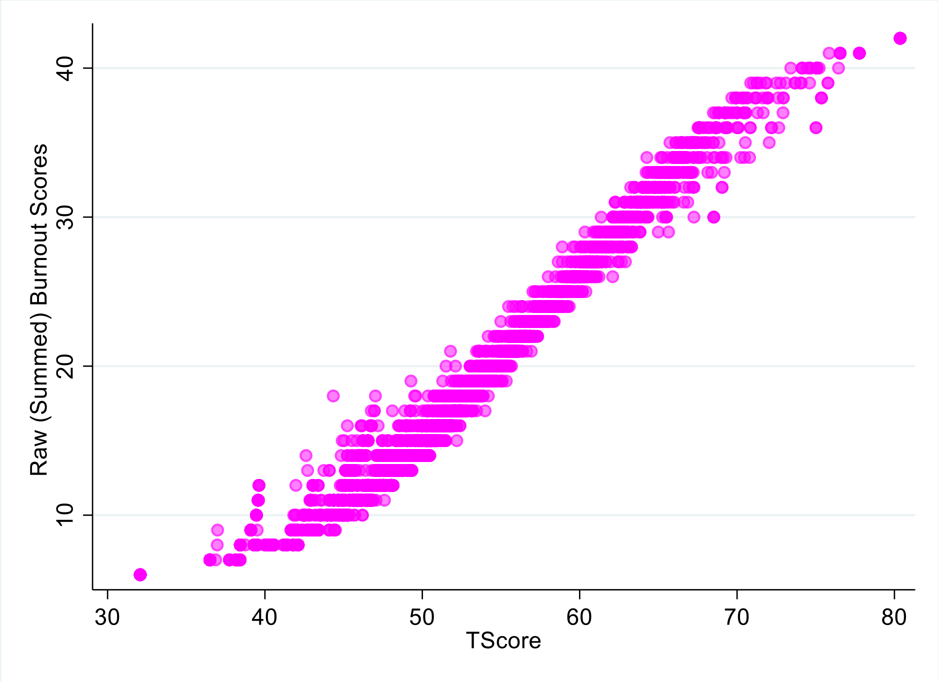
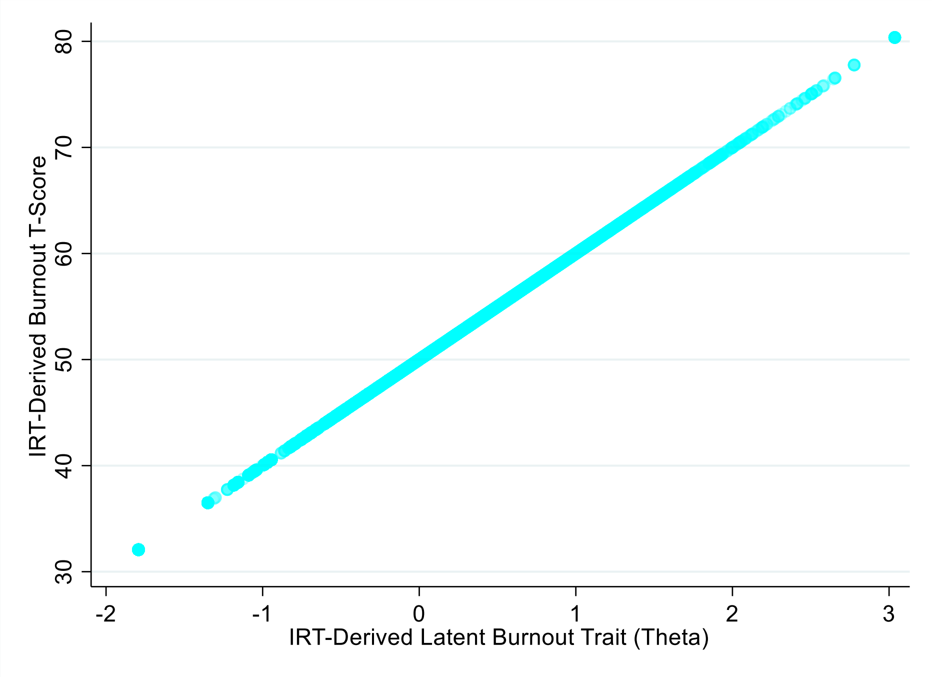


Table 10B. Scatterplot of IRT-Derived Burnout T-Scores against IRT-Derived Latent Burnout Trait (Theta)



APPENDIX TO SECTION E1

MOKKEN SCALE ANALYSIS OF 6-ITEM BURNOUT SCALE

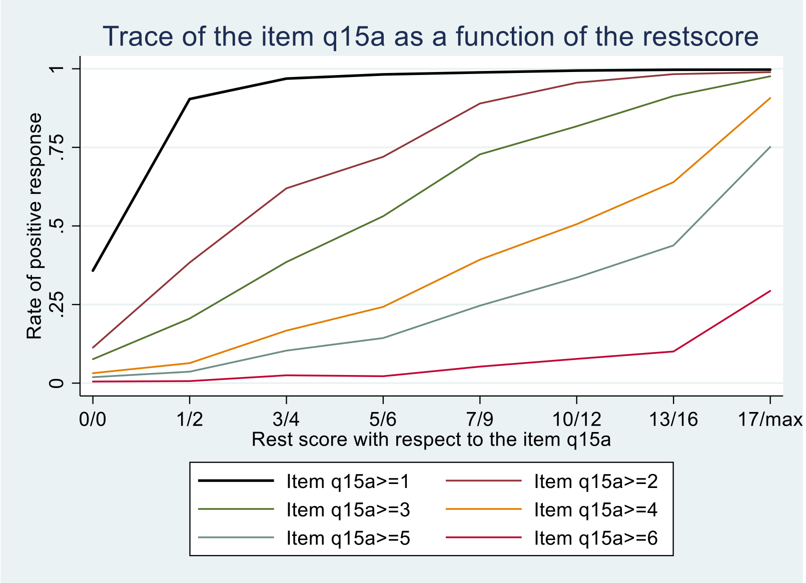
Stata output (`loevh` subroutine) is included on the following page.

Loveinger's *H* coefficient is a measure of an item's scalability, and is reported in the fourth column from the right in the top panel of output. The overall *H* scalability coefficient for the 6 items taken together is reported in the last row of the top panel of output. In the non-parametric item response model, monotone homogeneity model (MHM), item-level *H* and overall scale *H* should all be greater than 0.30, with values greater than 0.50 indicating strong scalability[40, 41] . All 6 burnout items demonstrate strong scalability. In a non-parametric MHM, such values would be considered indicative of unidimensionality.

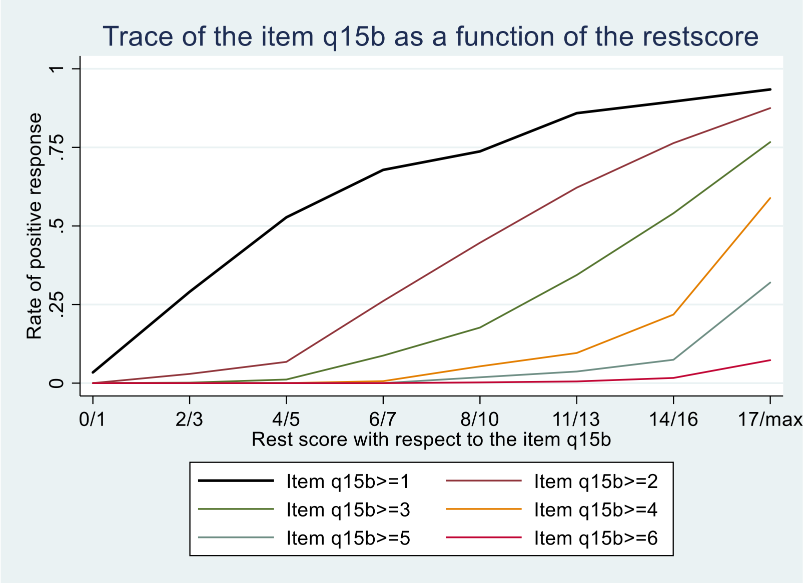
In the bottom panel, the number of departures from monotonicity (#vi) and statistically significant departures from monotonicity (#zsig) are reported. As shown, there were neither significant nor non-significant departures from the assumption of monotonicity among the 6 burnout items. Checks on monotonicity are further confirmed visually by the 6 trace plots, which graph the rate of a keyed response category by each item's rest score. Each item's rest score is simply the leave-out sum score – i.e. the total raw score minus the score on the item being evaluated. As can be seen, there is no visual evidence of non-monotonicity.

```
. loevh q15a-q15f, monotonicity(*)
                                  Observed    Expected                                      Number
                          Mean     Guttman     Guttman    Loevinger               H0: Hj<=0  of NS
Item          Obs         Score      errors      errors      H coeff     z-stat.     p-value    Hjk
---------------------------------------------------------------------------------------------------
q15a          5962      2.8380       29323    80560.11      0.63601     99.9193     0.00000      0
q15b          5962      1.4931       26758    75713.54      0.64659    105.9303     0.00000      0
q15c          5962      2.1798       24534    79655.81      0.69200    112.1380     0.00000      0
q15d          5962      1.4576       25418    76506.70      0.66777    109.2720     0.00000      0
q15e          5962      1.9277       26985    83608.48      0.67725    110.5815     0.00000      0
q15f          5962      0.8053       23990    59958.46      0.59989     90.9607     0.00000      0
---------------------------------------------------------------------------------------------------
Scale         5962                   78504     2.3e+05      0.65569    181.6061     0.00000

         Summary per item for check of monotonicity
         Minvi=0.030    Minsize=  597  Alpha=0.050

Items              #ac     #vi #vi/#ac     maxvi       sum   sum/#ac      zmax     #zsig    Crit
-------------------------------------------------------------------------------------------------
q15a               168       0                                                               -17
q15b                92       0                                                               -17
q15c               147       0                                                               -20
q15d               126       0                                                               -18
q15e               161       0                                                               -19
q15f               133       0                                                               -15  graph
-------------------------------------------------------------------------------------------------
Total             1654       0  0.0000    0.0000    0.0000    0.0000    0.0000         0
-------------------------------------------------------------------------------------------------
```

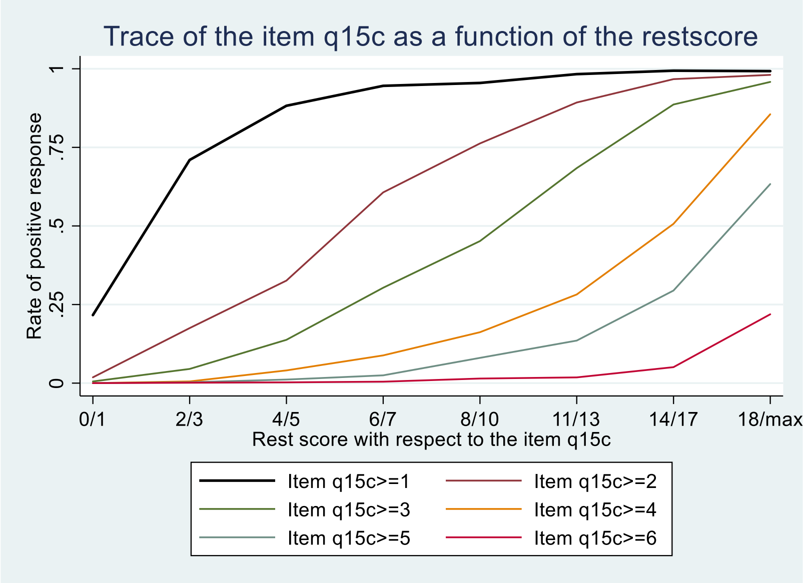
Trace Plot 1. Item Q15A



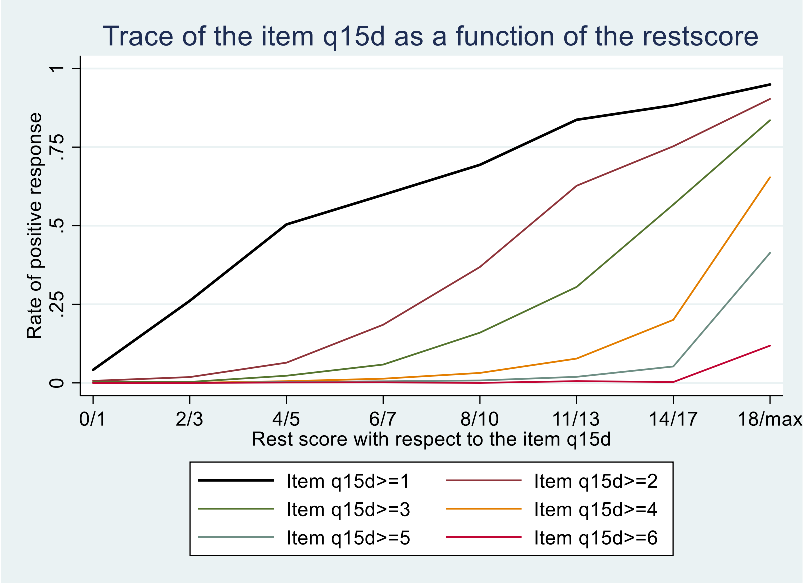
Trace Plot 2. Item Q15B



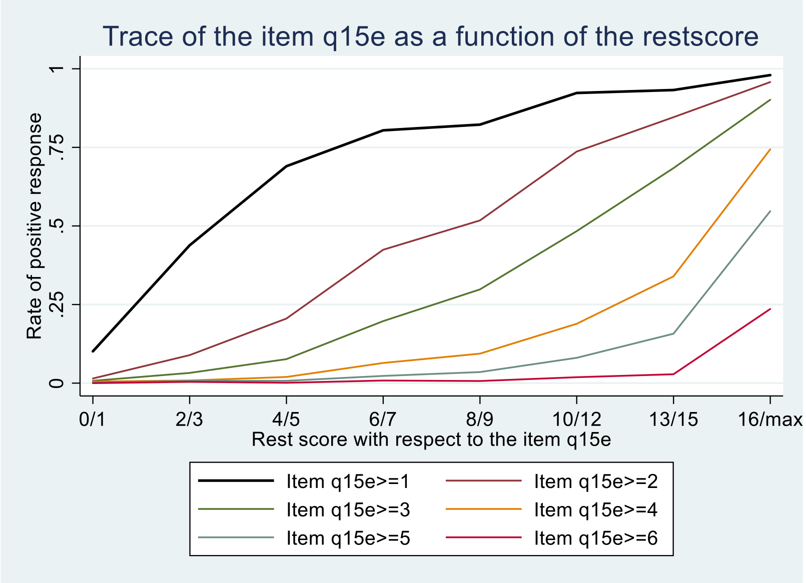
Trace Plot 3. Item Q15C



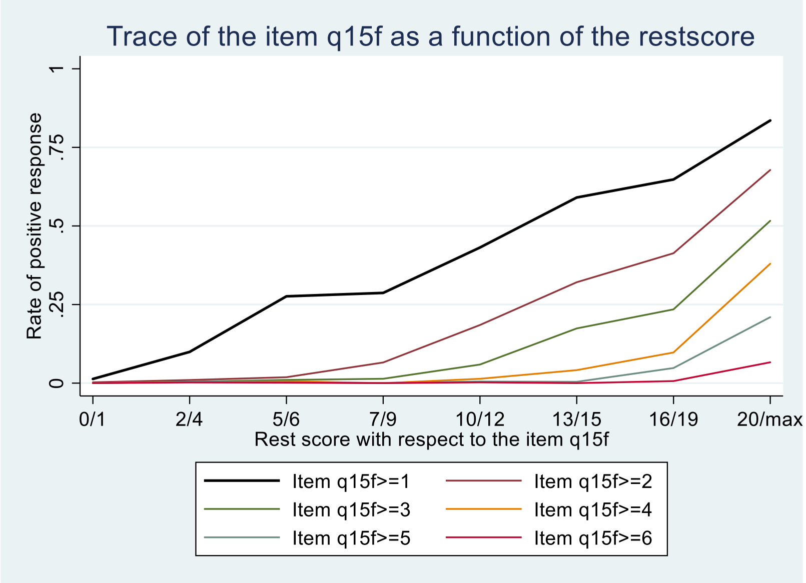
Trace Plot 4. Item Q15D



Trace Plot 5. Item Q15E



Trace Plot 6. Item Q15F



E.2. SECONDARY ENDPOINTS

This study will also assess secondary endpoints corresponding to resident well-being and the work environment (as characterized by the adapted Areas of Work Life discussed in §C.2). A domain score will be calculated as a weighed sum by Confirmatory Factor Analysis.

E.2.1. Resident Well-Being & Learning Environment Domain Scores

STUDY ENDPOINT	SOURCE	Item Name	Variable Name	Measure Type/Values
WELL-BEING – Thoughts of Attrition Percent	ABS	Q13	attrition	Continuous Numeric
WELL-BEING – Suicidality Percent	ABS	Q14	suicide	Continuous Numeric
WORKLOAD & JOB DEMANDS – Score	ABS		wjd_factor	Continuous Numeric
WORK-LIFE INTEGRATION – Score	ABS		worklife_factor	Continuous Numeric
EFFICIENCY & RESOURCES – Score	ABS		effres_factor	Continuous Numeric
RESIDENT CAMARADERIE – Score	ABS		rescam_factor	Continuous Numeric
FACULTY ENGAGEMENT – Score	ABS		faceng_factor	Continuous Numeric
ORGANIZATIONAL CULTURE & VALUES – Score	ABS		orgcv_factor	Continuous Numeric
CONTROL/FLEXIBILITY – Score	ABS		conflex_factor	Continuous Numeric
MEANING IN WORK – Score	ABS		miw_factor	Continuous Numeric
MISTREATMENT – Score	ABS		mistreat_factor	Continuous Numeric

E.2.2. Resident Education

STUDY ENDPOINT	SOURCE	Item Name	Variable Name	Measure Type/Values
EDUCATION OUTCOME – ABSITE Score (Raw)	ABS	ptot	ptot	Continuous Numeric
EDUCATION OUTCOME – ABSITE Score (Percentile by PGY Level)	ABS	percentile	percentile	Continuous Numeric
EDUCATION OUTCOME – ABSITE Score (Standard Score)	ABS	stdtot	stdtot	Continuous Numeric
EDUCATION OUTCOME – ABS Certifying Exam (CE) Pass Rate Percent	ABS	ce1pct	CE1	Continuous Numeric
EDUCATION OUTCOME – ABS Qualifying Exam (QE) Pass Rate Percent	ABS	qe1pct	QE1	Continuous Numeric
EDUCATION OUTCOME – ACGME Case Log Total Volume	ACGME	?	?	Continuous Numeric
EDUCATION OUTCOME – ACGME Case Log Category Volumes	ACGME	?	?	Continuous Numeric
EDUCATION OUTCOME – Resident Survey 80 hours	ACGME	?	?	Ordered Categorical
EDUCATION OUTCOME – Resident Survey Faculty & Staff Interested in Residency Education	ACGME	?	?	Ordered Categorical
EDUCATION OUTCOME – Resident Survey Faculty & Staff Create Environment of Inquiry	ACGME	?	?	Ordered Categorical
EDUCATION OUTCOME – Resident Survey Opportunity to Evaluate Program	ACGME	?	?	Ordered Categorical
EDUCATION OUTCOME – Resident Survey Satisfied that Program Uses Evaluations to Improve	ACGME	?	?	Ordered Categorical
EDUCATION OUTCOME – Resident Survey Instructed How to Manage Fatigue	ACGME	?	?	Ordered Categorical

EDUCATION OUTCOME – Resident Survey Satisfied with Opportunities for Scholarly Activities	ACGME	?	?	Ordered Categorical
EDUCATION OUTCOME – Resident Survey Appropriate Balance between Ed and Other Clinical Demands	ACGME	?	?	Ordered Categorical
EDUCATION OUTCOME – Resident Survey Education (Not) Compromised by Excessive Reliance on Non-Physician Obligations	ACGME	?	?	Ordered Categorical
EDUCATION OUTCOME – Resident Survey Supervisors Delegate Appropriately	ACGME	?	?	Ordered Categorical
EDUCATION OUTCOME – Resident Survey Provide a Way to Transition Care when Fatigued	ACGME	?	?	Ordered Categorical
EDUCATION OUTCOME – Resident Survey Satisfied with Process to Deal with Problems and Concerns	ACGME	?	?	Ordered Categorical
EDUCATION OUTCOME – Resident Survey Residents Can Raise Concerns without Fear	ACGME	?	?	Ordered Categorical
EDUCATION OUTCOME – Resident Survey Culture Reinforces Patient Safety Responsibility	ACGME	?	?	Ordered Categorical
EDUCATION OUTCOME – Resident Survey Information (Not) Lost during Shift Changes or Patient Transfers	ACGME	?	?	Ordered Categorical
EDUCATION OUTCOME – Resident Survey Effectively Work in Interprofessional Teams	ACGME	?	?	Ordered Categorical
EDUCATION OUTCOME – Resident Survey Program Provides Appropriately Graduated Supervision in the Operating Room	ACGME	?	?	Ordered Categorical
EDUCATION OUTCOME – Resident Survey Program Mandates Sufficient Experience as an Assistant in the OR before Allowing Residents to Act as Surgeon	ACGME	?	?	Ordered Categorical
EDUCATION OUTCOME – Resident Survey Program Encourages More Senior Residents to Act as Teaching Assistant to More Junior Residents in the OR	ACGME	?	?	Ordered Categorical

EDUCATION OUTCOME – Resident Survey Program Provides Appropriately Graduated Supervision in Non-Operative Patient Care	ACGME	?	?	Ordered Categorical
EDUCATION OUTCOME – Resident Survey Program Provides Sufficient Breadth and Depth of Experience/Confident Able to Practice Competently and Independently without Fellowship	ACGME	?	?	Ordered Categorical
EDUCATION OUTCOME – Faculty Survey Interest of Faculty & PD in Education	ACGME	?	?	Ordered Categorical
EDUCATION OUTCOME – Faculty Survey Worked on Scholarly Project with Residents/Fellows	ACGME	?	?	Ordered Categorical
EDUCATION OUTCOME – Faculty Survey Effectiveness of Graduating Residents/Fellows	ACGME	?	?	Ordered Categorical
EDUCATION OUTCOME – Faculty Survey Outcome Achievement of Graduating Residents/Fellows	ACGME	?	?	Ordered Categorical
EDUCATION OUTCOME – Faculty Survey Provide a Way to Transition Care when Fatigued	ACGME	?	?	Ordered Categorical
EDUCATION OUTCOME – Faculty Survey Residents/Fellows Workload Exceeds Capacity to Do the Work	ACGME	?	?	Ordered Categorical
EDUCATION OUTCOME – Faculty Survey Satisfied with process to Deal with Residents'/Fellows' Problems & Concerns	ACGME	?	?	Ordered Categorical
EDUCATION OUTCOME – Faculty Survey Prevent Excessive Reliance on Residents/Fellows to Fulfill Non-Physician Obligations	ACGME	?	?	Ordered Categorical
EDUCATION OUTCOME – Faculty Survey Information Not Lost During Shift Change or Patient Transfers	ACGME	?	?	Ordered Categorical
EDUCATION OUTCOME – Faculty Survey Culture Reinforces Patient Safety Responsibility	ACGME	?	?	Ordered Categorical

EDUCATION OUTCOME – Faculty Survey Residents/Fellows Communicate Effectively when Transferring Clinical Care	ACGME	?	?	Ordered Categorical
EDUCATION OUTCOME – Faculty Survey Residents/Fellows Effectively Work in Interprofessional Teams	ACGME	?	?	Ordered Categorical
EDUCATION OUTCOME – Faculty Survey Program Effective in Teaching Teamwork Skills	ACGME	?	?	Ordered Categorical

E.2.3. Patient Outcomes

STUDY ENDPOINT	SOURCE	NSQIP Item	Measure Type/Values
PATIENT OUTCOME – 30-Day Postoperative Death/Serious Morbidity	NSQIP	dsermorb	Dichotomous Categorical Yes, No
PATIENT OUTCOME – 30-Day Postoperative Cardiac Complications	NSQIP	compCard	Dichotomous Categorical Yes, No
PATIENT OUTCOME – 30-Day Postoperative Death	NSQIP	postcode	Dichotomous Categorical Yes, No
PATIENT OUTCOME – 30-Day Postoperative VTE	NSQIP	compDVT	Dichotomous Categorical Yes, No
PATIENT OUTCOME – 30-Day Postoperative Serious Morbidity	NSQIP	score1b	Dichotomous Categorical Yes, No
PATIENT OUTCOME – 30-Day Postoperative Pneumonia	NSQIP	compPneu	Dichotomous Categorical Yes, No
PATIENT OUTCOME – 30-Day Postoperative Readmission	NSQIP	readmission	Dichotomous Categorical Yes, No
PATIENT OUTCOME – 30-Day Postoperative Renal Failure	NSQIP	compRenal	Dichotomous Categorical Yes, No
PATIENT OUTCOME – 30-Day Postoperative Return to OR	NSQIP	returnor	Dichotomous Categorical Yes, No
PATIENT OUTCOME – 30-Day Postoperative Surgical Site Infection (SSI)	NSQIP	compSSI	Dichotomous Categorical Yes, No
PATIENT OUTCOME – 30-Day Postoperative Unplanned Reintubation	NSQIP	compTube	Dichotomous Categorical Yes, No
PATIENT OUTCOME – 30-Day Postoperative Urinary Tract Infection (UTI)	NSQIP	uti	Dichotomous Categorical Yes, No
PATIENT OUTCOME – Postoperative Ventilator>48 Hrs	NSQIP	compVent	Dichotomous Categorical Yes, No

PATIENT OUTCOME – Extended length of Stay (>75th Percentile)	NSQIP	los75	Dichotomous Categorical Yes, No

E.3. OTHER OUTCOMES

These outcomes are the individual items that comprise each domain score.

STUDY ENDPOINT	SOURCE	Item Name	Variable Name	Measure Type/Values
WORKLOAD & JOB DEMANDS - 80 Hour Workweek	ABS	Q8a	violate80	Ordered Categorical
WORKLOAD & JOB DEMANDS – 1 Day Off in 7	ABS	Q8b	violate1in7	Ordered Categorical
WORKLOAD & JOB DEMANDS – 1 in 3 Call	ABS	Q8c	violate1in3call	Ordered Categorical
WORKLOAD & JOB DEMANDS – Feel Pressured to Underreport Duty Hours	ABS		underreport	Ordered Categorical
WORK-LIFE INTEGRATION – Satisfaction with Time for Personal Life	ABS		personalsat	Ordered Categorical
WORK-LIFE INTEGRATION – Satisfaction with Ability to Maintain Healthy Habits	ABS		habitssat	Ordered Categorical
WORK-LIFE INTEGRATION – Satisfaction with Ability to Perform Routine Health Maintenance	ABS		maintenancesat	Ordered Categorical
EFFICIENCY & RESOURCES – Program Effectively Uses Support Staff	ABS	Q9b	apns	Ordered Categorical
EFFICIENCY & RESOURCES – Program Protects Educational Time	ABS	Q9c	protected	Ordered Categorical
EFFICIENCY & RESOURCES – Enough Time for Direct Patient Care after Completion of Administrative Tasks	ABS	Q9a	timeforcare	Ordered Categorical

EFFICIENCY & RESOURCES – Residency Coordinator is Helpful Resource	ABS		coordinator	Ordered Categorical
EFFICIENCY & RESOURCES – Adequate Computers, Workrooms, Materials	ABS		adequate	Ordered Categorical
EFFICIENCY & RESOURCES – RNs RESPECT and HELP	ABS		nurses	Ordered Categorical
CAMARADERIE – Belonging	ABS		belong	Ordered Categorical
CAMARADERIE – Work Appreciated by Co-Residents	ABS	Q9g	resappreciate	Ordered Categorical
CAMARADERIE – Residents Cooperate with One Another	ABS	Q9i	cooperate	Ordered Categorical
CAMARADERIE – Consider Co-Residents to be Among Closest Friends	ABS	Q9m	friends	Ordered Categorical
CAMARADERIE – Turn to Co-Residents for Support	ABS		support	Ordered Categorical
CAMARADERIE – Residents Step in to Help	ABS		stepin	Ordered Categorical
FACULTY ENGAGEMENT – Work Appreciated by Attendings	ABS	Q9h	attappreciate	Ordered Categorical
FACULTY ENGAGEMENT – Mentor within Surgery who Genuinely Cares	ABS	Q9k	mentor	Ordered Categorical
ORGANIZATIONAL CULTURE – Attendings are Good Role Models	ABS		rolemodels	Ordered Categorical
ORGANIZATIONAL CULTURE – Performance Evaluated Fairly	ABS		evalfair	Ordered Categorical
ORGANIZATIONAL CULTURE – Opportunities Distributed Fairly	ABS		oppsfair	Ordered Categorical

ORGANIZATIONAL CULTURE – Program Takes Wellness Seriously	ABS	Q9q	seriously	Ordered Categorical
ORGANIZATIONAL CULTURE – Program Emphasizes Learning from Adverse Events/Complications rather than Placing Blame	ABS	Q9o	learnblame	Ordered Categorical
ORGANIZATIONAL CULTURE – Program Helps Decompress/Debrief/Cope after Adverse Events	ABS	Q9n	debrief	Ordered Categorical
VOICE – Program Responsive to Resident Concerns	ABS	Q9l	responsive	Ordered Categorical
VOICE – Meaningful Input into Call and Vacation Schedules			meaningful	
VOICE – Comfortable Speaking Up			speakingup	
MEANING IN WORK – Appropriate Amount of Time in the OR	ABS	Q9d	ortime	Ordered Categorical
MEANING IN WORK – Appropriate Level of Autonomy in Patient Care & Clinical Decision-Making	ABS	Q9f	clinauto	Ordered Categorical
MEANING IN WORK – Appropriate Level of Operative Autonomy	ABS	Q9e	orauto	Ordered Categorical
MEANING IN WORK – Satisfaction with Decision to Become a Surgeon	ABS	Q10c	surgeonsat	Ordered Categorical
MISTREATMENT – Bullying	ABS	Q11a-j	bully	Continuous Numeric
MISTREATMENT – Sexual Harassment	ABS	Q16a-f	sexharass	Ordered Categorical
MISTREATMENT –Discrimination Based on Childcare	ABS	Q18a-g	dischild	Ordered Categorical

MISTREATMENT – Gender/Gender Identity/Sexual Orientation Discrimination	ABS	Q18a-g	disgender	Ordered Categorical
MISTREATMENT – Race/Ethnicity/Religious Discrimination	ABS	Q20a-f	disrace	Ordered Categorical

E.4. OTHER MEASURES

Other measures that may be used in statistical analyses may include, but are not limited to:

Resident Characteristic	SOURCE	Item Name	Variable Name	Measure Type/Values
PGY level	ABS	level	pgy	Ordered Categorical PGY1, PGY2, PGY3, PGY4, PGY5
Gender	ABS	Gender	gender	Unordered Categorical Male, Female
Clinically active	ABS	Q1	clinactive	Unordered Categorical Yes, No, Unsure
Relationship status	ABS	Q2	marital	Unordered Categorical Married, Not Married but in a Relationship, Not Married and Not in a Relationship, Divorced/Separated, Widowed
Number of children	ABS	Q3	children	Ordered Categorical 1, 2, 3, 4, 5+
Pregnant/Adopting/Expecting	ABS	Q4	children2	Unordered Categorical Yes, No, Unsure
Race	ABS	Q5a-Q5g	nativeamerican, asian, black, pacific, white, raceother, raceprefer	Dichotomous: American Indian/Alaska Native Asian Black/African American Native Hawaiian/Other Pacific Islander White Other

				Prefer Not to Say
Ethnicity	ABS	Q6	hispanic	Unordered Categorical Hispanic or Latino, Not Hispanic or Latino, Prefer Not to Say
LGBTQ	ABS	Q7a-7g	straight, gay, bisexual, orientother, trans, otherident, identprefer	Dichotomous: Straight Gay/Lesbian Bisexual Other Orientation Transgender Other Gender Prefer Not to Say

Program Characteristic	SOURCE	Item Name	Variable Name	Measure Type/Values
Proportion Residency Program International Medical Graduates (IMG)	ABS	propimg	PCTIMG	Continuous Numeric
Proportion Residency Program Male	ABS	propmale	PCTMALE	Continuous Numeric
Program Size	ABS	progsize	PGMSIZE	Continuous Numeric
Program Type	ABS	progtype	PGMTYPE	Categorical Academic, Community-Based, Military
Geographic Region	ABS	Geo	Geo	Categorical Northeast, South, Midwest, Southwest, West
Current Number of PGY 5 Residents	ABS	PGY5_N	PGY5_N	Continuous Numeric
Number of ABSITE Examinees	ABS	RES_N	RES_N	Continuous Numeric
Proportion Faculty Female	AAMC	percfemalefac	pctfemfac	Continuous Numeric

Chair Gender		chair_gender	ch_gender	Unordered Categorical: Male, Female
Program Director Gender		pd_gender	pd_gender	Unordered Categorical: Male, Female

Patient Characteristic	SOURCE	Item Name	Variable Name	Measure Type/Values
Treatment Group	SECOND	tx_arm	TREAT	Dichotomous Categorical Intervention, Control
Age	NSQIP	age	age	Continuous Numeric
Age Group	NSQIP	agegroup	AGECAT	Ordered Categorical <65, 65-74, 75-84, 85+
ASA Class	NSQIP	asaclas_c	ASACLASS45	Categorical Class 1, Class 2, Class 3, Classes 4&5
Ascites	NSQIP	ascites	ASCITES	Dichotomous Categorical Yes, No
Bleeding Disorders	NSQIP	bleeddis	BLEED	Dichotomous Categorical Yes, No
BMI Class (Ref: Normal)	NSQIP	bmi_class	BMICLASS	Categorical Normal, Class 1 Obese, Class 2 Obese, Class 3 Obese, Overweight, Underweight
Congestive Heart Failure	NSQIP	hxchf	CHF	Dichotomous Categorical Yes, No
CPT Linear Risk – Cardiac	NSQIP	Cardlin	Cardlin	Continuous Numeric
CPT Linear Risk – Death	NSQIP	postcodelin	postcodelin	Continuous Numeric
CPT Linear Risk – Death/Ser Morb	NSQIP	dsermorblin	dsermorblin	Continuous Numeric
CPT Linear Risk – DVT	NSQIP	DVTlin	DVTlin	Continuous Numeric
CPT Linear Risk – Morbidity	NSQIP	score1blin	score1blin	Continuous Numeric
CPT Linear Risk – Pneumonia	NSQIP	pneulin	Pneulin	Continuous Numeric
CPT Linear Risk – Readmission	NSQIP	readmissionlin	readmissionlin	Continuous Numeric
CPT Linear Risk – Renal Failure	NSQIP	Renallin	Renallin	Continuous Numeric
CPT Linear Risk – Return OR	NSQIP	returnorlin	returnorlin	Continuous Numeric
CPT Linear Risk – SSI	NSQIP	SSIlin	SSIlin	Continuous Numeric
CPT Linear Risk – Unplanned Intubation	NSQIP	Tubelin	Tubelin	Continuous Numeric
CPT Linear Risk – UTI	NSQIP	UTIlin	UTIlin	Continuous Numeric
CPT Linear Risk – Ventilator >48hrs	NSQIP	Ventlin	Ventlin	Continuous Numeric
Current Smoker	NSQIP	smoke	SMOKE	Dichotomous Categorical Yes, No

Diabetes	NSQIP	diabetes	DIABETES	Categorical No Diabetes, Insulin, Oral
Preoperative Renal Failure (Ref: No)	NSQIP	renafail	PRRENAL	Dichotomous Categorical Yes, No
Preoperative Sepsis	NSQIP	prsepis	PRSEPIS	Dichotomous Categorical No, SIRS, Sepsis, Septic Shock
Preoperative SGOT>40 (Ref: ≤40)	NSQIP	sgot1	SGOT40	Dichotomous Categorical ≤40, >40
Preoperative Sodium	NSQIP	prsodm	** varies across mod	Categorical ** Thresholds vary across models
Preoperative WBC	NSQIP	prwbc	** varies across mod	Categorical ** Thresholds vary across models
Race	NSQIP	race_class	RACECAT	Categorical White, American Indian/Native Alaskan, Black/African American, Native Hawaiian/Pacific Islander, Asian, Unknown
Sex Male	NSQIP	sex	SEXMALE	Dichotomous Categorical Female, Male
Steroid Use	NSQIP	steroid	STEROID	Dichotomous Categorical Yes, No
Transfer Status	NSQIP	transt_c	TRANS	Categorical Admit from Home, Transfer from Acute Care, Transfer from Chronic Care, Transfer from Other Facility, Transfer from Outside ED
Transfusion	NSQIP	transfuse	TRANSFUSION	Dichotomous Categorical Yes, No
Ventilator Dependent	NSQIP	ventilat	VENTDEP	Dichotomous Categorical Yes, No
Weight Loss >10%	NSQIP	wtloss	WTLOSS	Dichotomous Categorical Yes, No
Work RVU	NSQIP	workrvu	workrvu	Continuous Numeric
Wound Class (Ref: Clean)	NSQIP	wndclas	WOUND	Categorical Clean, Clean/Contaminated, Contaminated, Dirty/Infected

*Other patient characteristics from ACS NSQIP may be utilized as well.

Hospital Characteristic	SOURCE	Item Name	Variable Name	Measure Type/Values
Bed Size (Ref: ≤100 Beds)	AHA	BDTOT	BEDCAT1	Ordered Categorical ≤100 Bed, 101-300 Beds, 301-500 Beds, 500+ Beds
Total Annual Discharge Volume	AHA	ADMTOT	ADMTOT	Continuous Numeric
Joint Commission Accreditation	AHA	MAPP1	JCAHO	Dichotomous Categorical Yes, No

ACS Approved Cancer Program	AHA	MAPP2	ACSCANCER	Dichotomous Categorical Yes, No
COTH Member	AHA	MAPP8	COTH	Dichotomous Categorical Yes, No
Nurse-to-Bed Staffing Ratio	AHA	FTERN, FTELPN, BDTOT	NURSBED	Continuous Numeric (FTERN + FTELPN)/BDTOT
Geographic Region (Ref: Midwest)	AHA	MSTATE	REGION	Categorical West, South, Northeast, Southeast
CMS Case Mix Index	CMS	cmiv30	cmiv30	Continuous Numeric
Resident-to-Bed Ratio	CMS	residenttobed	residenttobed	Continuous Numeric
Ownership	AHA	CNTRL	TYPCNTRL1	Categorical Not-For-Profit, For-Profit, Government
Total Surgical Volume	AHA	SUROPTOT	SUROPTOT	Continuous Numeric
Total Inpatient Surgical Volume	AHA	SUROPIP	SUROPIP	Continuous Numeric
Total Outpatient Surgical Volume	AHA	SUROPOP	SUROPOP	Continuous Numeric
Percent Surgical Volume – Inpatient	AHA	SUROPIP, SUROPTOT	PCTSUROPIP=SUROPIP/SUROPTOT	Continuous Numeric
Percent Surgical Volume – Outpatient	AHA	SUROPOP, SUROPTOT	PCTSUROPOP=SUROPOP/SUROPTOT	Continuous Numeric
Level 1 Trauma Designation	AHA	TRAUML90	LEVEL1	Dichotomous Categorical Yes, No
ANCC Nursing Magnet Status	ANCC	Coded from ANCC website	NURSEMAGNET	Dichotomous Categorical Yes, No
Medical/Surgical Intensive Care Unit Beds	AHA	MSICBD	MSICBD	Continuous Numeric
Transplant Hospital	AHA	OTBONHOS, HARTHOS, KDNYHOS, LIVRHOS, LUNGHOS, TISUHOS, OTOTHHOS	TRANSPLANT (Bone, Heart, Kidney, Liver, Lung, Tissue OR Other Transplant)	Dichotomous Categorical Yes, No
Rural	AHA	CBSATYPE	RURAL (CBSATYPE Micro)	Dichotomous Categorical Yes, No

Baseline Quality (Death/Serious Morbidity Odds Ratio)	NSQIP			Continuous Numeric

*****Other hospital structural characteristics from the AHA Annual Survey and CMS Impact File may be utilized as well.

E.5. DATA TRANSFORMATIONS

E.5.1. Categorical Variables

Many variables in the SECOND Trial Resident Survey and Program Directors' Survey are categorical. Several are measured using 5-point Likert-type scales. Frequencies of all variables in their untransformed state will be tabulated and examined. However, categorical variables may be recoded for reasons that may include:

- Re-coding and collapsing categories to avoid small cells, which we define in this study as having ≤5 observations

- Dichotomizing to facilitate analyses – in particular, the interpretability of associations

E.5.2. Continuous Variables

The distribution of natural and derived continuous variables (for example, program-level rates of resident respondents) in the main analyses will be characterized by its mean, standard deviation, median, skewness, and kurtosis, among other statistics. Continuous variables may be transformed for reasons that may include:

- Discretization due to non-normality or to facilitate interpretation of associations
- Log-transformation or other transformation to address non-normality, particularly in the context of conforming to statistical assumptions underlying a particular modeling technique

E.6. PROCEDURES FOR MISSING DATA

E.6.1. Complete Case Analysis

Primary analyses will rely on complete case analysis in which only observations for which complete data on all variables for a given model exist. The drawback of this approach is that if data are not missing *at random*, parameter estimates may be biased. A second limitation of this approach is that, depending on the volume of incomplete data, analytic sample attrition due to missing data may attenuate statistical power. However, because the SECOND Trial Resident Survey is optional to residents, and because many survey items may be socially sensitive, missing data are likely to be missing *not at random*. Although patterns of missing data will be explored and multiple imputation methods for addressing missingness will also be explored, complete case analysis will be the first approach due to its simplicity. Limitations of this approach will be disclosed in reports of analyses relying on complete case analyses.

E.6.2. Multiple Imputation

Data from the SECOND Trial Resident Survey will be inspected and patterns of missing data will be explored and characterized to ascertain whether or not data may be missing at random. If the data suggest it would be reasonable to do so, multiple imputation by chained equations (MICE) will be used to fill missing values.[31]

SECTION F

STATISTICAL METHODS AND PLANNED ANALYSES

SECTION OVERVIEW

F.1. ANALYSIS PRINCIPLES

F.1.1. Intent-to-Treat

An intent-to-treat (ITT) approach will be adopted for the main analysis. In this approach, all programs and residents will be included in the analysis “as assigned,” irrespective of whether programs adopted (Intervention Arm) or did not adopt (Control) any organizational changes or measures during the study period to address resident wellbeing and/or the learning environment. The ITT estimate of the intervention effect will be biased if exposure is independent of (un)observable characteristics/features of programs and residents that would be correlated with both exposure and outcomes (i.e. programs and residents are randomized to exposure); if attrition and missing data are at random and if endpoints are measured without bias.

The ITT approach is particularly appropriate for this particular study due to the nature of the intervention. The intervention is the provision (or ‘offer’) of information: (a) program-specific feedback on resident wellness outcomes, as well as on the learning environment; and (b) a multicomponent Wellness Toolkit of various strategies for program-level organizational responses to promote and/or improve resident wellness and the learning environment within a program. The intervention was designed as an “offer of information” because the Principal Investigators believe different programs have different problem areas – i.e. areas that need improvement with respect to resident wellness and the learning environment. Thus, rational improvement should be targeted to program-specific problem areas. Therefore, programs may (or may not) adopt different measures to address different problems.

F.1.2. Per-Protocol

Sensitivity analyses will include a per-protocol approach in which we may control for the level of adherence (see §D.12.1) to study arm protocols and/or exclude non-adherent entities from analyses.

F.1.3. As-Treated

Sensitivity analyses will include an as-treated approach in which we investigate the effect of increased wellness activities/resources (see §D.12.1) on patient and resident outcomes, regardless of study arm assignment.

F.2. DESCRIPTIVE STATISTICS

There remains controversy regarding the appropriateness of statistical testing to evaluate balance of baseline characteristics across study arms following randomization.[43-45] In accordance with CONSORT 2010 guidelines,[45] we will assess baseline characteristics of programs and residents across both study arms, but we will refrain from statistical testing for differences across arms, because any differences across study arms will be known to have been due to chance.

F.3. STATISTICAL EVALUATION PLAN

F.3.1. Analysis of Primary Outcome

The primary study hypothesis is that:

- Residents in programs randomized to Intervention will have lower mean burnout scores compared to residents in programs randomized to Control

F.3.1.1. Unadjusted Analyses

Under randomization of programs to treatment arms, treatment should be orthogonal to observed and unobserved confounders. An unadjusted comparison of mean burnout scores across study arms would provide an intent-to-treat estimate of the effect of study assignment on resident burnout. We may carry this out using a cluster-corrected Student's t-test to compare mean burnout scores across study arms, and/or we may estimate a simple hierarchical linear regression of the form

$$y_{ij} = \alpha_0 + \beta_0 TREAT_j + \gamma_1 Q_{2j} + \gamma_2 Q_{3j} + \gamma_3 Q_{4j} + \gamma_4 Q_{5j} + \zeta_j + \epsilon_{ij}$$

where y_{ij} is the burnout T-score of resident *i* in program j*,* $TREAT_j$ is the study arm assignment of program j, and Q_{2j}-Q_{5j} are dummy variables indicating program j*'s* randomization stratum (program quartile ranking based on 2019 ABSITE burnout scores, plus a fifth stratum for programs with no 2019 ABSITE data). In this model, under randomization of programs to study arms, the coefficient β_0 is the unadjusted population-averaged intent-to-treat estimate of the effect of assignment to intervention on mean resident burnout. Other parameters in the model are the constant (α_0), coefficients on program-level randomization strata (γ_1-γ_m), program-level random intercepts (ζ_j) and individual-level errors (ϵ_{ij}).

F.3.1.2. Adjusted Analyses

Covariate adjustment in evaluations of randomized trials is generally done for purposes of increasing statistical efficiency by reducing variance within treatment groups.[46, 47] Covariate adjustment is also sometimes also undertaken on the basis of chance imbalance of baseline characteristics between study groups. Although programs will be randomized to treatment arms in this study, we anticipate considerable variation in the number of ABSITE respondents per program. In the 2019 ABSITE data, the average program had 23 resident respondents (mean=23.11, SD=11.15; coefficient of variation=48%), but the range spanned from a minimum of 3 residents per program to a maximum of 59 residents per program. Such dispersion in cluster sizes may lead to imbalance of resident characteristics

Thus, our analytic approach to evaluating our primary hypothesis will be to estimate a 2-level hierarchical linear model of the general form:

$$y_{ij} = \alpha_0 + \beta_0 TREAT_j + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \cdots + \beta_k x_{kij} + \gamma_1 z_{1j} + \cdots + \gamma_m z_{mj} + \zeta_j + \epsilon_{ij}$$

where y_{ij} is the burnout T-score of resident *i* in program j*,* $TREAT_j$ is the study arm assignment of program j*,* and x_{1ij}*...*x_{kij} are resident-level covariates and z_{1j}*...*z_{mj} are program-level covariates to be adjusted for, including sampling stratification variables. In this model, under randomization of programs to study arms, the coefficient β_0 is the adjusted population-averaged intent-to-treat estimate of the effect of assignment to intervention on mean resident burnout. A coefficient $\beta_0 < 0$ that is statistically significant with a one-tailed p<0.0443 provides evidence supporting the effectiveness of the intervention over usual care. Other parameters in the model are the constant (α_0), coefficients on other resident-level covariates (β_1-β_k), coefficients on program-level covariates (γ_1-γ_m), program-level random intercepts (ζ_j) and individual-level errors (ϵ_{ij}).

Covariates *may* include variables listed in §E.3.: resident gender; resident relationship status; resident race; program type; program size; and randomization variables such as the quartile of program j ranked by 2019 program-level mean burnout scores. This list of covariates is subject to change.

Covariates will not include any other resident-level outcome or behavioral variable(s) that may be affected by the intervention as doing so may introduce endogeneity. Thus, for example, measures of career satisfaction, duty hour adherence, mistreatment, discrimination/harassment etc. will not be considered as controls in the evaluation of burnout.

Any structural measures – i.e., measures of program characteristics, policy, and/or programs related to burnout will not be controlled for in the primary analysis as doing so might over-control and attenuate the intent-to-treat estimate of the treatment assignment effect.

We will examine the mean, median, skewness and kurtosis values of the dependent variable to assess the normality of the dependent variable. If there is evidence of non-normality in the dependent variable, we may consider an appropriate transformation of the dependent variable;[46] however, generalized linear mixed models are flexible and can accommodate alternative distributional assumptions regarding y_{ij} without altering the original scale of y_{ij}.

In accordance with the decision of the Principal Investigators, estimates from the adjusted analyses of the primary outcome will be reported as the primary findings of this study, although both adjusted and unadjusted analyses will be conducted.

F.3.1.3. Program-Level Analyses

To assess the robustness of our findings across alternative modeling approaches, we will also evaluate the effect of assignment to Intervention on aggregate, program-level mean burnout T-scores. For these analyses, we will construct mean burnout T-scores for each program in the study. We will then regress mean program burnout T-scores on a dichotomous variable indicating study arm assignment, as well as randomization strata as follows:

$$MeanBurnout_j = \alpha + \beta_0 TREAT_j + \gamma_1 Q_{2j} + \gamma_2 Q_{3j} + \gamma_3 Q_{4j} + \upsilon_j$$

where $MeanBurnout_j$ is the mean burnout score in program j, $TREAT_j$ is the study arm assignment of program j, and $Q_{2j} - Q_{4j}$ are randomization strata (assuming quartiles of previous year's mean burnout scores are used for stratification, with Q_{1j} omitted as the reference category). Here, β_0 is the intent-to-treat estimate of the effect of assignment to Intervention on mean program burnout rates, and we hypothesize $\beta_0 < 0$.

We may additionally estimate adjusted program-level regressions by incorporating program-level covariates into the model above.

We anticipate estimating this program-level model using ordinary least squares (OLS) regression, pending assessment of program-level mean burnout T-scores for normality. In the 2019 ABSITE data, program-mean burnout T-scores had a skewness of -0.558 and kurtosis of 3.587. Shapiro-Wilk, Shapiro-Francia, and skewness-kurtosis tests for normality were significant. Decisions regarding modeling and transformation of the dependent variable will depend on the normality of program-level mean burnout scores in the actual study data, as well as considerations of statistical assumptions and interpretability of results.

F.3.2. Analysis of Secondary Outcomes

As discussed in §C.2 and §E.2, a number of secondary endpoints may be evaluated pertaining to resident wellbeing, resident education, and patient outcomes.

F.3.2.1. Secondary Resident Outcomes

F.3.2.1.A. Shanafelt Areas of Work-Life

As discussed in §C.2, we developed scores to measure seven domains of work-life balance:

- Workload & Job Demands
- Efficiency & Resources
- Organizational Culture
- Meaning in Work
- Resident Camaraderie
- Faculty Engagement
- Voice
- Work-Life Integration
- Mistreatment

Each domain score is a continuous variable. To evaluate the effect of randomization on resident scores within each of these seven domains, we will estimate a 2-level hierarchical random intercept regression model of the general form:

$$y_{ij} = \alpha_0 + \beta_0 TREAT_j + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \cdots + \beta_k x_{kij} + \gamma_1 z_{1j} + \cdots + \gamma_m z_{mj} + \zeta_j + \epsilon_{ij}$$

where y_{ij} is the domain score of resident *i* in program *j* in a particular domain. Separate models will be estimated for each domain score. In each model, outcomes will be regressed on study arm assignment of program *j* ($TREAT_j$), and $x_{1ij} \ldots x_{kij}$ are resident-level covariates and $z_{1j} \ldots z_{mj}$ are program-level covariates to be adjusted for, including sampling stratification variables.

F.3.2.1.B. Component Items of Areas of Work-Life

As discussed in §C.2, responses to individual survey items comprising each of the 9 domains may themselves be analyzed as secondary outcomes. Most of these component items are measured on a Likert-type scale and are thus ordered categorical variables. Several analytic approaches may be taken in the evaluation of these component outcomes.

A two-level hierarchical ordered logistic regression model can be estimated to model the log-odds that a resident will select a response category above a level *s* as a function of study arm assignment, resident characteristics, program characteristics, and program intercepts. The general form of these models would be:

$$log\left(\frac{Pr\left(y_{ij} > s\middle|TREAT_j, x, z, \zeta_j\right)}{(1 - \left(Pr\left(y_{ij} > s\middle|TREAT_j, x, z, \zeta_j\right)\right)}\right) =$$
$$\beta_0 TREAT_j + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \cdots + \beta_k x_{kij} + \gamma_1 z_{1j} + \cdots + \gamma_m z_{mj} + \zeta_j - \kappa_s$$

where y_{ij} is the response of resident *i* in program j; x_{1ij}...x_{kij} are resident-level covariates and z_{1j}...z_{mj} are program-level covariates to be adjusted for, including sampling stratification variables, ζ_j are program intercepts, and κ_s are category cutoff parameters. Such models can, in theory, be estimated using generalized linear latent and mixed models (`gllamm`) in *Stata*. In practice, such models are computationally intensive and are unlikely to be practical. Moreover, ordered logistic regression rests on the assumption of proportional odds: that is, the association between a covariate and the outcome is constant across categories. There is no ready way to test or address this assumption following estimation of a hierarchical ordered logistic regression in *Stata*.

A more practical approach will be to estimate a non-hierarchical generalized ordered logistic regression as above, with robust clustered standard errors to account for program-level clustering. Generalized ordered logistic regression estimates a single set of parameters for variables that meet the proportional odds assumption. For those variables that violate this assumption, generalized ordered logistic regression estimates separate parameters for each level.

Interpreting and reporting the results of generalized ordered logistic regression models can be cumbersome. For an outcome with 5 response categories, up to 5-1=4 parameters may be estimated for each covariate that violates the proportional odds assumption. For outcomes with 7 response categories (e.g. mistreatment survey items), up to 6 parameters may be estimated for each covariate that violates the proportional odds assumption. Thus, a more practical approach may be to dichotomize these ordered categorical variables. Dichotomization will permit estimation of hierarchical logistic regression equations of the general form:

$$log\left(\frac{Pr\left(y_{ij} = 1\middle|TREAT_j, x, z, \zeta_j\right)}{(1 - \left(Pr\left(y_{ij} = 1\middle|TREAT_j, x, z, \zeta_j\right)\right)}\right) =$$

$$\beta_0 TREAT_j + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \cdots + \beta_k x_{kij} + \gamma_1 z_{1j} + \cdots + \gamma_m z_{mj} + \zeta_j$$

where y_{ij} is the response of resident *i* in program j; x_{1ij}...x_{kij} are resident-level covariates and z_{1j}...z_{mj} are program-level covariates to be adjusted for, including sampling stratification variables and ζ_j are program intercepts.

Alternatively, or perhaps additionally, non-hierarchical logistic regression equations may be estimated with robust standard errors that account for non-independence of observations within programs.

F.3.2.1.C. Education Outcomes

As discussed in §C.2.2, responses to individual survey we will also evaluate the effect of randomization to intervention on various resident educational outcomes, including ABSITE scores, ABS Qualifying

Examination (QE) scores, ABS Certifying Examination (CE) scores, ACGME case log data, and ACGME annual resident and faculty surveys.

Some of these outcomes, such as test scores, will be treated as continuous variables and will be evaluated using hierarchical linear models of the general form described in §F.3.2.1.A. Dichotomous categorical outcomes will be evaluated using hierarchical logistic models of the general form described in §F.3.2.1.B.

F.3.2.2. Secondary Patient Outcomes

As discussed in §C.2.3, we will evaluate the effect of program participation in Intervention on various postoperative outcomes of patients undergoing surgery in hospitals associated with those programs compared to patients in hospitals associated with programs that did not participate in the Intervention. This evaluation will only be carried out among programs associated with hospitals that participate in the ACS NSQIP program.

Patient outcomes are dichotomous variables indicating whether a patient experienced each of the following postoperative complications:

- 30 day postoperative death or serious morbidity
- 30 day postoperative death
- 30 day postoperative morbidity (NSQIP composite)
- 30-Day Postoperative Surgical Site Infection (SSI)
- 30-Day Postoperative Respiratory Complications (composite of pneumonia, intraoperative or postoperative unplanned intubation, pulmonary embolism, ventilation ≥48 hours)
- 30-Day Postoperative Pneumonia
- 30-Day Postoperative Urinary Tract Infection
- 30-Day Postoperative Renal Failure
- 30-Day Postoperative Venous Thromboembolism (deep vein thrombosis and/or pulmonary embolism)
- 30-Day Postoperative Stroke
- 30-Day Postoperative Myocardial Infarction
- 30-Day Intra- and Postoperative Transfusion
- 30-Day Postoperative Sepsis/Septic Shock
- 30-Day Postoperative Unplanned Return-to-Operating Room
- 30-Day Postoperative Unplanned Readmission
- Non-Home Discharge
- Failure to Rescue
- 30-Day Postoperative Cardiac Complications

Each outcome will be modeled separately using hierarchical logistic regression models of the general form

$$log\left(\frac{Pr\left(complication_{ijr}=1\middle|TREAT_{ijk},x,z,\zeta_j^{(2)},\zeta_r^{(3)}\right)}{(1-\left(Pr\left(complication_{ijr}=1\middle|TREAT_{ijk},x,z,\zeta_j^{(2)},\zeta_r^{(3)}\right)\right)}\right)=$$

$$\beta_0 TREAT_j + \beta_1 x_{1ijr} + \beta_2 x_{2ijr} + \cdots + \beta_k x_{kijr} + \gamma_1 z_{1jr} + \cdots + \gamma_m z_{mjr} + \zeta_j^{(2)} + \zeta_r^{(3)}$$

where $complication_{ij}$ is the outcome of patient *i* in hospital *j* that is associated with program *r*; $x_{1ijr} \ldots x_{kir}$ are patient covariates, $z_{1jr} \ldots z_{mjr}$ are hospital-level covariates to be adjusted for, $z_{1r} \ldots z_{wr}$ are program-level covariates including sampling stratification variables, $\zeta_{jr}^{(2)}$ are hospital-level intercepts and $\zeta_{jr}^{(3)}$ are program-level covariates. Patient-level covariates for risk-adjustment will be based on outcome-specific risk-adjustment models developed by NSQIP for use in their Semi-Annual Reports (SAR).

F.3.3.1. Pre-Post with Control Analyses

As an additional perspective, we may explore analyses that take baseline rates of programs in each arm into account. In these analyses, rather than identifying the intervention on the basis of a post-only comparison between Intervention and Control, we examine whether *improvement* in outcomes was greater among programs randomized to Intervention relative to improvement in outcomes among programs randomized to Control. These analyses may be executed by regressing pre-post differences in outcomes on study arm assignment and controls for sampling strata as well as other covariates at the discretion of Investigators and analysts, as appropriate to the data. We do not take this approach as our main analytic strategy because the Intervention under investigation is largely untested. Prospective power calculations would require two assumptions: (a) the rate of secular improvement in the absence of the intervention; and (b) the effectiveness of the intervention or its components to estimate the expected difference in mean rates within the Intervention group.

F.3.3.2. Planned Subgroup Analyses

We may conduct subgroup analyses to explore the possibility of heterogeneous intervention effects across theoretically/policy/socially-relevant subgroups of programs and/or residents, depending on the level of analysis. In program-level analyses, we may conduct subgroup analyses to investigate whether there was any intervention modification by characteristics that may include: baseline level of program activity regarding resident wellness or learning environment, geographic region, program size, baseline levels of resident burnout and/or mistreatment, gender balance or racial balance. In resident-level analyses, we may conduct subgroup analyses to investigate whether there were differential effects of Intervention for female versus male residents; for non-white versus white residents; or for burned-out versus non-burned out residents. We may also examine whether there was any intervention modification by region, residency year, or baseline wellbeing, for example. All subgroup analyses will be executed by means of regression models that include an interaction term between study arm assignment and indicators of subgroups of interest.[48] Statistical tests on interaction coefficients will provide evidence for the presence or absence of effect modification. Subgroup analyses will not be executed by subgroup stratification.

F.3.3.4. Local Average Treatment Effects

In a pragmatic trials such as this, there is little control of experimental conditions. In this trial, there is no condition in Control that restricts programs randomized to Control from engaging in any new activity or continuing any former activity related to promoting resident wellness or improving the learning environment. Although programs randomized to Intervention are offered information and other facilitative information to promote resident wellness and improve the learning environment, there is likewise no condition that requires programs randomized to Intervention to review the information provided, much less adopt any recommended measures. In more traditional trial settings where adherence to study conditions can be even loosely defined, one can address issues of compliance heterogeneity by estimating local average treatment effects, which captures the effect of the intervention on compliers.[49]

In this trial, we may adopt this approach and use instrumental variables (IV) to estimate the effect of intervention on outcomes among the subset of programs that acted in response to the intervention. Specifically, we will use study arm assignment to instrument implementation of a program organizational measure for promoting resident wellness and/or improving the learning environment that was not previously in place prior to the study intervention. We will then regress outcomes on instrumented implementation to obtain the average intervention effect among programs that would implement program improvement if offered the intervention. This analytic approach only uses variation in implementation that was due to random exposure of intervention to predict outcomes, and is thus free endogenous implementation due to self-selection on unobservables. We anticipate study arm assignment to serve as a valid instrument because: (a) study arm was randomly assigned; (b) study arm is assumed to have no effect on outcomes except through implementation of an intervention (exclusion criterion) and (c) study arm is predictive of implementing an intervention.

We may use two-stage least squares (TSLS) IV models for program-level estimation of local average treatment effects. First-stage F statistics will be examined to assess strength of the instrument in predicting implementation of program improvement interventions. We will not proceed if treatment assignment turns out to be a weak instrument.[50]

F.3.4. General Procedures

F.3.4.1. Descriptive Statistics

Dependent and independent variables in all models will be described using appropriate statistics which may include simple frequencies, mean, standard deviation, median, skewness, and kurtosis.

F.3.4.2. Bivariate Statistics

The preceding subsections have largely dealt with multivariable regression models. As a preliminary step in modeling efforts, we will also investigate simple bivariate relationships between outcomes and explanatory variables of interest using appropriate bivariate statistics which may include but are not limited to t-tests, one-way analysis of variance, chi-squared or Fisher's exact tests.

F.3.4.3. Model Specification and Evaluation

The preceding sections describe our general analytic approach. In practice, we aim to specify models that are appropriate to the demands of the data; therefore, we reserve the right to amend and/or alter our statistical methods and approach based model diagnostics and/or new techniques that come to our attention. Different models will also be explored to assess the robustness of our qualitative findings to alternative specifications.

F.3.4.4. Multiple Comparisons

In the evaluation of secondary outcomes, we anticipate multiple comparisons within a family of outcomes where we consider families of outcomes to be groups of similar outcome measures within a domain (e.g.

resident education outcomes). To control for familywise error rate in conducting multiple hypothesis tests, we will adjust the overall study p-value (p=0.0444, see §F.4) using the Bonferroni correction.

F.4. STATISTICAL SIGNIFICANCE LEVEL

The overall level of statistical significance for this study will be set at $p<0.05$. One interim analysis is planned for the midpoint of this study. Assuming an O'Brien-Fleming alpha-spending function, the overall alpha=0.05 will be adjusted (LANDEMETS, Stata module) to $\text{alpha}_{\text{interim}}=0.0056$ for the interim analysis and $\text{alpha}_{\text{final}}=0.0444$ for the final analysis.

In analyses where there are multiple comparisons – for example, as in the case of multiple tests of a single hypothesis using different indicators for the same underlying construct in separate models, p-values will be adjusted using Bonferroni's method or another similar method to control for the family-wise error rate.

F.5. PLANNED INTERIM ANALYSIS

F.5.1. Interim Analysis Overview

A planned Interim Analysis will be undertaken at the midpoint of the Study Period. The purpose of the Interim Analysis is to assess whether there are any excess resident safety/wellness risks with respect to burnout that would raise ethical concerns about trial continuation.

The Interim Analysis will assess whether there are statistically significant differences in resident burnout scores across study arms (§F.3.1.2).

Results of the Interim Analysis will be reported to the DSMB (§D.12). Based on results of the Interim Analysis, the DSMB is charged with deciding whether the Trial should be terminated early or if the Trial can continue per protocol. Results of the Interim Analysis will not be used to modify Trial design or protocol.

F.5.2. Schedule of Interim Analysis

The planned Interim Analysis is scheduled for April 2021. This is approximately midpoint through the SECOND Trial Study period, as the first SECOND Trial Resident Survey after dissemination of the Intervention will occur in January 2021.

ACS NSQIP data for the subset of SECOND sites that participate is scheduled to be delivered to the Study Team in January 2021. Because of ACS NSQIP data collection procedures which allow participating hospitals 120 days after the date of a procedure to complete data collection and entry.

A confidential Interim Report will be circulated to DSMB members in May 2021. A DSMB conference call is scheduled for May 2021.

F.5.3. Trial Stopping Rules

The DSMB will review the results of the analysis and will decide whether any statistically significant differences in resident burnout scores across arms are clinically significant. The DSMB will have the responsibility to recommend trial continuation or early termination. Early termination will be based on two considerations: (1) finding statistically significant differences in burnout scores between Control and Intervention arms, and (2) DSMB finding the statistically significant differences ***are also clinically significant*** – i.e., that the magnitude of differences pose real welfare advantages/disadvantages to residents. If both of these conditions hold, and the DSMB votes to recommend early termination of the trial, then the trial will be terminated following the DSMB meeting to discuss the Interim Analysis.

F.5.4. Interim Analysis Statistical Methods

We expect data for the Interim Analysis to come from the 2021 SECOND Trial Resident Survey. Resident burnout will be measured as previously described in §E.1. We will follow statistical methods that have already described for the analysis of resident burnout, which are outlined in §F.3.1. As discussed in §F.4., the level of statistical significance for the Interim Analysis will be set at $p_{Interim}<0.0056$ to maintain an overall significance level of $p<0.05$ for the SECOND Trial as a whole.

F.6. DATA MANAGEMENT AND ANALYSIS SOFTWARE

Analytic software to be used in the SECOND Trial may include, but is not limited to:

- Stata MP Version 15
- SAS
- Microsoft Excel
- MPlus

SECTION G

LIMITATIONS

G.1. LIMITATIONS OF GENERAL STUDY DESIGN

- Participants in both Intervention and Control groups likely already have programs in place that address various aspects of resident wellbeing. As a pragmatic trial, we are not able or willing to request that participating programs deviate from what they have already implemented. Thus, the SECOND Trial intervention is properly seen as an offer of information (feedback reports and information on alternative interventions) on top of what may already be in place. Lack of a clean baseline and control group will attenuate any measurable effect of intervention towards zero.

G.2. LIMITATIONS OF SURVEY INSTRUMENTS AND ADMINISTRATION

- Data on mistreatment obtained through our survey instruments may be subject to a number of respondent biases that have been documented and discussed in various literature.[51]

- In the past, our survey has been administered at the conclusion of the annual American Board of Surgery In-Training Examination (ABSITE), which is a 5-hour examination administered each January. Responses may be biased in either direction by either exam-related distress/fatigue or post-exam elation.

- Furthermore, various items in the survey have been administered in previous years. Thus, there is a chance that senior residents may have become sensitized to the survey instrument.

G.3. LIMITATIONS OF INTERVENTION

- Components of the intervention were developed by studying what selected programs had already implemented. Components that may have worked for a specific program may not be feasible and/or effective in another program due to differences in organizational and/or institutional environments.

- The intervention is only an offer of information – both information on a program’s own performance, as well as information on possible modes of action. However, no financial resources are given to programs to adopt and implement any action. Action may be differentially costly depending not only on the specific component to be implemented, but also depending on a program’s level of discretionary resources. Thus, failure to find any appreciable effect of randomization to intervention on resident burnout may not necessarily be due to the usefulness of the information itself, but due to insurmountable barriers to implementation.

G.4. LIMITATIONS OF STATISTICAL METHODS

- Our current implementation of item response theory (IRT) graded response model (GRM) does not account for the hierarchical nature of the ABSITE resident data, in which residents are clustered within residency programs. In other words, IRT GRM parameters do not account for any clustering at the program level. Hierarchical IRT models have been developed, but are not readily implemented within IRT subroutines in standard software packages such as *Stata*. Hierarchical IRT models for ordered, polytomous variables are even less developed. We acknowledge our use of non-hierarchical IRT GRM in our current work. However, we note that our final models in which T-Scores are used as the outcome variable *do* account for the hierarchical nature of the data (§D.8). It is unclear what the implications are of ignoring program-specific effects in estimating the discrimination and location parameters in IRT GRM models. At this time, we surmise that finding a way to implement hierarchical IRT GRM models may entail greater costs than benefit.

- Given the large number of secondary outcomes, it may be unlikely that we will find any statistically significant effect of the intervention on secondary outcomes after controlling for familywise error rates.

SECTION H

STUDY IMPLEMENTATION TIMELINE

In order to successfully implement the SECOND Trial, numerous forms of correspondence with programs need to occur. Correspondence serves as a way to disseminate information to participating hospitals, as well as, for hospitals to provide us with information needed to assure adherence to protocol.

Date	Implementation Activity
1.2.19	First Email for Program Tours Sent
4.3.19	PD Survey Email
4.4.19	Informational Webinar for PDs
4.10.19	Informational Webinar for PDs
4.25.19	APDS Town Hall
6.12.19	PD Survey Reminder Email
6.26.19	Enrollment Email
8.5.19	Enrollment Reminder Email (PDs and Chairs)
8.13.19	Enrollment Reminder Email (DIOs)
8.28.19	Personal Contacts to Unenrolled Programs
8.30.19	Enrollment Deadline Extended
9.4.19	Clarification of Eligibility to NYS Programs
9.30.19	Enrollment Closed
11.2019	Randomization
12.2019	Program Specific Report Delivery
1.2020	Wellness Toolkit Delivery to Intervention
1.2020	SECOND Trial Resident Survey
4.2020	PD Survey
6.2020	Program Specific Report Delivery
1.2021	SECOND Trial Resident Survey
4.2021	PD Survey
5.2021	Interim Report Completed
5.2021	DSMB Conference Call
6.2021	Program Specific Report Delivery
1.2022	SECOND Trial Resident Survey
4.2022	PD Survey
6.2022	Program Specific Report Delivery
1.2023	SECOND Trial Resident Survey
1.2023	Wellness Toolkit Delivery to Control
4.2023	PD Survey
5.2023	Final Report Completed
5.2023	DSMB Conference Call
6.2023	Program Specific Report Delivery

REFERENCES

1. Fnais N, Soobiah C, Chen MH, et al. Harassment and discrimination in medical training: a systematic review and meta-analysis. *Acad Med*. May 2014;89(5):817-27. doi:10.1097/ACM.0000000000000200
2. Hu Y-Y, Ellis RJ, Hewitt B, et al. Discrimination, abuse, harassment, and burnout in surgical residency training. *The New England Journal of Medicine*. 2019;381(18):1741-1752.
3. Lebares CC, Guvva EV, Ascher NL, O'Sullivan PS, Harris HW, Epel ES. Burnout and Stress Among US Surgery Residents: Psychological Distress and Resilience. *J Am Coll Surg*. Jan 2018;226(1):80-90. doi:10.1016/j.jamcollsurg.2017.10.010
4. McManus IC, Winder BC, Gordon D. The causal links between stress and burnout in a longitudinal study of UK doctors. *Lancet (London, England)*. Jun 15 2002;359(9323):2089-90.
5. Sullivan MC, Yeo H, Roman SA, et al. Surgical residency and attrition: defining the individual and programmatic factors predictive of trainee losses. *J Am Coll Surg*. Mar 2013;216(3):461-71. doi:10.1016/j.jamcollsurg.2012.11.005
6. Gifford E, Galante J, Kaji AH, et al. Factors associated with general surgery residents' desire to leave residency programs: a multi-institutional study. *JAMA Surg*. Sep 2014;149(9):948-53. doi:10.1001/jamasurg.2014.935
7. Shanafelt TD, Balch CM, Dyrbye L, et al. Special report: suicidal ideation among American surgeons. *Arch Surg*. Jan 2011;146(1):54-62. doi:10.1001/archsurg.2010.292
8. Bilimoria KY, Chung JW, Hedges LV, et al. National cluster-randomized trial of duty-hour flexibility in surgical training. *The New England Journal of Medicine*. 2016;374(8):713-727.
9. Rajaram R, Chung JW, Jones AT, et al. Association of the 2011 ACGME resident duty hour reform with general surgery patient outcomes and with resident examination performance. *JAMA*. Dec 10 2014;312(22):2374-84. doi:10.1001/jama.2014.15277
10. Bilimoria KY, Chung JW, Hedges LV, et al. Development of the Flexibility in Duty Hour Requirements for Surgical Trainees (FIRST) Trial Protocol: A National Cluster-Randomized Trial of Resident Duty Hour Policies. *JAMA Surg*. Mar 2016;151(3):273-81. doi:10.1001/jamasurg.2015.4990
11. Blay E, Jr., Hewitt DB, Chung JW, et al. Association Between Flexible Duty Hour Policies and General Surgery Resident Examination Performance: A Flexibility in Duty Hour Requirements for Surgical Trainees (FIRST) Trial Analysis. *J Am Coll Surg*. Feb 2017;224(2):137-142. doi:10.1016/j.jamcollsurg.2016.10.042
12. Bilimoria KY, Hoyt DB, Lewis F. Making the Case for Investigating Flexibility in Duty Hour Limits for Surgical Residents. *JAMA Surg*. Jun 2015;150(6):503-4. doi:10.1001/jamasurg.2015.0239
13. Dahlke AR, Quinn CM, Chung JW, Bilimoria KY. Surgical Residents' Work Hours and Well-Being in Year 2 of the FIRST Trial. *N Engl J Med*. Jul 13 2017;377(2):192-194. doi:10.1056/NEJMc1703812
14. Bilimoria KY, Hoyt DB, Lewis FR. Surgical Resident Duty Hours. *N Engl J Med*. Jun 16 2016;374(24):2402-3. doi:10.1056/NEJMc1604659
15. Summary of Changes to ACGME Common Program Requirements Section VI. . 2017;
16. Ellis RJ HA, Hewitt DB, Engelhardt KE, Yang AD, Merkow RP, Bilimoria KY, Hu YY. A comprehensive national survey on thoughts of leaving residency, alternative career paths, and reasons for staying in general surgery training. *Am J Surg*. 2020;219(2):227-232.
17. Rotenstein LS, Torre M, Ramos MA, et al. Prevalence of Burnout Among Physicians: A Systematic Review. *JAMA*. Sep 18 2018;320(11):1131-1150. doi:10.1001/jama.2018.12777

18. Dyrbye LN, Burke SE, Hardeman RR, et al. Association of Clinical Specialty With Symptoms of Burnout and Career Choice Regret Among US Resident Physicians. *JAMA*. Sep 18 2018;320(11):1114-1130. doi:10.1001/jama.2018.12615
19. Maslach CJ, Susan E.; Leiter, Michael P. *Maslach Burnout Inventory Manual Fourth Edition*. Mind Garden, Inc.; 2016.
20. Shanafelt TD, Noseworthy JH. Executive Leadership and Physician Well-being: Nine Organizational Strategies to Promote Engagement and Reduce Burnout. *Mayo Clin Proc*. Jan 2017;92(1):129-146. doi:10.1016/j.mayocp.2016.10.004
21. Ellis RJ HD, Hu YY, Johnson JK, Merkow RP, Yang AD, Potts JR 3rd, Hoyt DB, Buyske J, Bilimoria KY. An Empirical National Assessment of the Learning Environment and Factors Associated With Program Culture. *Ann Surg*. 2019;270(4):585-592. doi:10.1097/SLA.0000000000003545
22. Cohen ME ea. Optimizing ACS NSQIP modeling for evaluation of surgical quality and risk: patient risk adjustment, procedure mix adjustment, shrinkage adjustment, and surgical focus. *J Am Coll Surg*. 2013;217(2):336-46 e1.
23. Shiloach M ea. Toward robust information: data quality and inter-rater reliability in the American College of Surgeons National Surgical Quality Improvement Program. *J Am Coll Surg*. 2010;210(1):6-16.
24. Merkow RP, Hall BL, Cohen ME, et al. Validity and feasibility of the American College of Surgeons colectomy composite outcome quality measure. *Ann Surg*. Mar 2013;257(3):483-9. doi:10.1097/SLA.0b013e318273bf17
25. Fowler JH, Christakis NA. Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the Framingham Heart Study. *Bmj*. Dec 4 2008;337:a2338. doi:10.1136/bmj.a2338
26. Petitta L, Jiang L, Hartel CEJ. Emotional contagion and burnout among nurses and doctors: Do joy and anger from different sources of stakeholders matter? *Stress Health*. Oct 2017;33(4):358-369. doi:10.1002/smi.2724
27. Reise S. Item Response Theory. In: L CR, O LS, eds. *The Encyclopedia of Clinical Psychology*. 1 ed. John Wiley & Sons, Inc.; 2015.
28. Reise SP, Henson JM. A discussion of modern versus traditional psychometrics as applied to personality assessment scales. *J Pers Assess*. Oct 2003;81(2):93-103.
29. Reise SP, Ainsworth AT, Haviland MG. Item response theory - Fundamentals, applications, and promise in psychological research. *Curr Dir Psychol Sci*. Apr 2005;14(2):95-101.
30. Thomas ML. The Value of Item Response Theory in Clinical Assessment: A Review. *Assessment*. Sep 2011;18(3):291-307.
31. Harwell MR, Gatti GG. Rescaling ordinal data to interval data in educational research. *Rev Educ Res*. Spr 2001;71(1):105-131.
32. Edelen MO, Reeve BB. Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of life research*. 2007;16(S1):5-18. doi:10.1007/s11136-007-9198-0
33. Edelen MO, Reeve BB. Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research*. 2007;16:5-18.
34. Reise SP, Moore TM, Haviland MG. Bifactor Models and Rotations: Exploring the Extent to Which Multidimensional Data Yield Univocal Scale Scores. *J Pers Assess*. 2010;92(6):544-559. doi:10.1080/00223891.2010.496477
35. Jin-Shei L, Paul KC, David C. Factor Analysis Techniques for Assessing Sufficient Unidimensionality of Cancer Related Fatigue. *Quality of life research*. 2006;15(7):1179-1190. doi:10.1007/s11136-006-0060-6

36. Reise SP, Scheines R, Widaman KF, Haviland MG. Multidimensionality and Structural Coefficient Bias in Structural Equation Modeling: A Bifactor Perspective. *Educational and Psychological Measurement*. 2013;73(1):5-26. doi:10.1177/0013164412449831

37. Bonifay WE, Reise SP, Scheines R, Meijer RR. When Are Multidimensional Data Unidimensional Enough for Structural Equation Modeling? An Evaluation of the DETECT Multidimensionality Index. 2015;

38. Rodriguez A, Reise SP, Haviland MG. Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological methods*. 2016;21(2):137-150. doi:10.1037/met0000045

39. Reise SP, Haviland MG. Item response theory and the measurement of clinical change. *J Pers Assess*. Jun 2005;84(3):228-238.

40. Stochl J, Jones PB, Croudace TJ. Mokken scale analysis of mental health and well-being questionnaire item responses: a non-parametric IRT method in empirical research for applied health researchers. *BMC medical research methodology*. 2012;12(1):74-74. doi:10.1186/1471-2288-12-74

41. Sijtsma K, van der Ark LA. A tutorial on how to do a Mokken scale analysis on your test and questionnaire data. *British journal of mathematical & statistical psychology*. 2017;70(1):137-158.

42. Baker FB. *The basics of item response theory*. 2nd ed.. ed. ERIC Clearinghouse on Assessment and Evaluation; 2001.

43. Austin PC, Manca A, Zwarenstein M, Juurlink DN, Stanbrook MB. A substantial and confusing variation exists in handling of baseline covariates in randomized controlled trials: a review of trials published in leading medical journals. *Journal of clinical epidemiology*. Feb 2010;63(2):142-53. doi:10.1016/j.jclinepi.2009.06.002

44. de Boer MR, Waterlander WE, Kuijper LD, Steenhuis IH, Twisk JW. Testing for baseline differences in randomized controlled trials: an unhealthy research behavior that is hard to eradicate. *The international journal of behavioral nutrition and physical activity*. Jan 24 2015;12:4. doi:10.1186/s12966-015-0162-z

45. Moher D, Hopewell S, Schulz KF, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *Bmj*. Mar 23 2010;340:c869. doi:10.1136/bmj.c869

46. Manning WG, Mullahy J. Estimating log models: to transform or not to transform? *Journal of health economics*. Jul 2001;20(4):461-94.

47. Kahan BC, Jairath V, Dore CJ, Morris TP. The risks and rewards of covariate adjustment in randomized trials: an assessment of 12 outcomes from 8 studies. *Trials*. Apr 23 2014;15:139. doi:10.1186/1745-6215-15-139

48. Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM. Statistics in medicine--reporting of subgroup analyses in clinical trials. *N Engl J Med*. Nov 22 2007;357(21):2189-94. doi:10.1056/NEJMsr077003

49. Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. *J Am Stat Assoc*. Jun 1996;91(434):444-455.

50. Bound J, Jaeger DA, Baker RM. Problems with Instrumental Variables Estimation When the Correlation between the Instruments and the Endogenous Explanatory Variable Is Weak. *J Am Stat Assoc*. Jun 1995;90(430):443-450.

51. Lewis TT, Cogburn CD, Williams DR. Self-Reported Experiences of Discrimination and Health: Scientific Advances, Ongoing Controversies, and Emerging Issues. *Annu Rev Clin Psycho*. 2015;11:407-440.