

OFFICIAL TITLE: A PILOT STUDY OF COGNITIVE BEHAVIORAL THERAPY –
INSOMNIA (CBT-I) AMONG PROFESSIONAL FIREFIGHTERS

NCT NUMBER: NCT05087121

DOCUMENT DATE: 6/15/2022

STUDY PROTOCOL AND STATISTICAL ANALYSIS PLAN

STUDY PROTOCOL AND STATISTICAL ANALYSIS PLAN

Method

Participants

Participants were recruited from fire departments in the Pacific Northwestern U.S. through fliers, email, and snowball recruitment. Requirements for inclusion in the study were: 1) English-speaking, 2) full-time professional firefighter, 3) at least 18 years of age, 4) a global Pittsburgh Sleep Quality Index score ≥ 5 (Grandner et al., 2006) and, 5) be willing and able to participate in all study activities including pre-, post-, and 3-month follow up assessments and 6-weeks of the CBT-I-F intervention. Participant exclusion criteria were: 1) endorsement of active suicidal ideation, 2) endorsement of prior involvement in formal CBT-I interventions, and 3) unwillingness to give written informed consent. Twenty-three firefighters met inclusion criteria and were randomly assigned using a ratio of 1.5:1 (CBT-I-F: waitlist control group), via stratified randomization for gender, to one of two groups: CBT-I-F or waitlist control group (WLC). Randomization was stratified for gender based on the low percentage of female firefighters across the U.S. (Evarts & Stein, 2020). Participants in the WLC condition were offered the opportunity to participate in the treatment following the study.

While feasibility and acceptability studies generally do not require the inclusion of an a priori power analysis, a test examining projected sample size was conducted to determine potential sample size needed given the inclusion of hypotheses regarding preliminary outcomes. Previous literature examining the effects of CBT-I on sleep quality (Okajima et al., 2011), suggest medium-to-large effect sizes ($f = 0.24-0.41$), when using a power level of 0.80, and an

alpha of 0.05. There does not appear to be prior research citing correlations between repeated measures for the PSQI in CBT-I interventions. Prior studies have suggested using test-retest correlations for this value (Koffel et al., 2015), however, doing so would overinflate the decay in correlation for repeated measures within an intervention study. Conversely, measures administered to an intervention group do generally maintain some consistency over time relative to the measures administered to the comparison group and for these reasons, the default value of 0.5 was used for correlation among repeated measures. Power analyses were conducted via G*Power (Faul et al., 2009). To obtain a power level of .80, an alpha of .05, and a medium effect size ($f = 0.24$), a total of 30 participants would be needed (18 allocated to CBT-I-F, and 12 allocated to WLC) for a 2x3 design with a correlation among repeated measures of 0.5. This would produce an 81% chance of correctly rejecting the null hypothesis of the interaction. To obtain a power level of .80, an alpha of .05, and a large effect size ($f = 0.41$), a total of 12 participants would be needed (7.2 allocated to CBT-I-F, and 4.8 allocated to WLC) for a 2x3 design with a correlation among repeated measures of 0.5. This would produce an 83% chance of correctly rejecting the null hypothesis of the interaction. The final sample size that was obtained contained 23 participants ranging in age from 24 to 55 ($M = 37.74$, $SD = 10.25$), and 100% were male. The CBT-I-F condition contained 13 participants, and the WLC condition contained 9 participants. Additional demographics are provided in Table 1.

Measures

Demographics

Demographic questions included age, race/ethnicity, gender, education, as well as occupation-specific items such as full-time employment status, standard shift schedule (e.g., 24/48 vs. 48/96), current position/rank within department, and years employed as a firefighter.

Self-Report Measures

Pittsburgh Sleep Quality Index (PSQI). The PSQI is a self-report questionnaire designed to assess sleep quality and sleep disfunction throughout the previous month (Buysse et al., 1989). Global scores range between 0-21 with higher scores indicating greater sleep dysfunction, and are comprised of seven subcomponents. The PSQI has demonstrated good internal consistency ($\alpha = 0.70-0.83$; Buysse et al., 1989; Mollayeva et al., 2016), and has a high test-retest reliability for both short durations (2-day interval; $r = 0.90$) and longer durations (45.6 ± 18 -day interval; $r = 0.86$; Backhaus et al., 2002). Due to the drastic variability in sleep quality that firefighters often experience between on-duty and off-duty settings, two separate PSQI questionnaires will be included in this study; one reflecting sleep quality only “at work” and one reflecting sleep quality only while “at home”. In the present study, the PSQI demonstrated unacceptable-to-acceptable internal consistency (pre-PSQI(Work) $\alpha = 0.41$; post-PSQI(Work) $\alpha = 0.24$; 3-mo. follow-up-PSQI(Work) $\alpha = 0.40$; pre-PSQI(Home) $\alpha = 0.66$; post-PSQI(Home) $\alpha = 0.56$; 3-mo. follow-up-PSQI(Home) $\alpha = 0.42$).

Consensus Sleep Diary (CSD). The CSD is a daily self-report measure used to track time in/out of bed (TIB; time elapsed from when the lights are turned out until the patient gets out of bed in the morning), total wake time (TWT; total time awake during the period allotted for sleep, time to fall asleep + time awake in middle of night + time awake in bed in morning) sleep onset latency (SOL; time it takes to fall asleep), total sleep time (TST; total duration of sleep), sleep efficiency (SE; percentage of time in bed that is spent asleep), and wakefulness after sleep onset (WASO; the time awake in the middle of the night; Carney et al., 2012; Manber & Carney, 2015). The CSD has demonstrated sensitivity to treatment, shown significant correlations in

expected directions with measures of insomnia ($r = -.43-.70$), and has a high compliance rate (99.8%; Maich et al., 2018).

Depression Anxiety Stress Scales 21 (DASS-21). The DASS-21 is a 21-item self-report measure of depression, anxiety, and stress, and is a short form of the 42-item Depression Anxiety Stress Scale (DASS; Lovibond & Lovibond, 1995). Items are responded to on a 4-point Likert-type scale ranging from 0 (*Did not apply to me at all*) to 3 (*Applied to me very much or most of the time*). Scores are totaled for each of the three, 7-item subscales and multiplied by two to reflect full-length DASS scores (Henry & Crawford, 2005). Higher scores indicate a greater severity of depression, anxiety, and stress. The DASS-21 has demonstrated good internal consistency for each of the subscales (Depression $\alpha = 0.88-0.91$; Anxiety $\alpha = 0.80-0.87$; Stress $\alpha = 0.84-0.91$; Henry & Crawford, 2005; Sinclair et al., 2012). In the present study, the DASS-21 demonstrated unacceptable-to-good internal consistency for the subscales of Depression (pre-DASS-21 Depression $\alpha = 0.88$; post-DASS-21 Depression $\alpha = 0.78$; 3-mo. follow-up-DASS-21 Depression $\alpha = 0.84$), Anxiety (pre-DASS-21 Anxiety $\alpha = 0.63$; post-DASS-21 Anxiety $\alpha = 0.095$; 3-mo. follow-up-DASS-21 Anxiety $\alpha = 0.51$), and Stress (pre-DASS-21 Stress $\alpha = 0.63$; post-DASS-21 Stress $\alpha = 0.79$; 3-mo. follow-up-DASS-21 Stress $\alpha = 0.85$).

PTSD Checklist for DSM-5 (PCL-5). The PCL-5 is a 20-item self-report measure of posttraumatic stress disorder (PTSD) with items based directly on diagnostic criteria defined by the DSM-5 (Blevins et al., 2015). Items are responded to on a 5-point Likert-type scale ranging between 0 (*Not at all*) to 4 (*Extremely*) with higher scores indicating greater severity of PTSD symptoms. Items are summed to create a total score ranging between 0-80 and scores greater than 33 indicate probable PTSD (Bovin et al., 2016). The PCL-5 has demonstrated high internal consistency ($\alpha = .94$) and test-retest reliability ($r = .82$; Blevins et al., 2015). In the present study,

the PCL-5 demonstrated good-to-excellent internal consistency (pre-PCL-5 $\alpha = 0.82$; post-PCL-5 $\alpha = 0.80$; 3-mo. follow-up-PCL-5 $\alpha = 0.91$).

Medical Outcomes Study 12-item Short-Form Survey (SF12v2). The SF12v2 is a 12-item self-report measure of general health (Ware et al., 2002; Ware & Sherbourne, 1992). The SF12v2 is a brief version of the longer 116-item core Medical Outcomes Survey and contains two component summary measures; the Physical Component Summary (SF_PCS) and the Mental Component Summary (SF_MCS). Higher scores indicate more favorable states of health. The SF12v2 has demonstrated acceptable internal consistency ($\alpha = 0.76$ - 0.90 ; Jenkinson et al., 1993; Ware, 2000). In the present study, the SF12v2 demonstrated unacceptable internal consistency (pre-SF12v2 $\alpha = 0.17$; post-SF12v2 $\alpha = 0.25$; 3-mo. follow-up-SF12v2 $\alpha = 0.23$).

McMaster Family Assessment Device – General Functioning Scale (FAD-GFS). The FAD-GFS is a 12-item self-report subscale of the larger 60-item FAD measure, that assesses overall health and pathology of the family (Byles et al., 1988; Miller et al., 1985). Items on the FAD-GFS are responded to on a 4-point Likert-type scale ranging between 1 (*Strongly Agree*) and 4 (*Strongly Disagree*). Higher total scores indicate more problematic family functioning. The FAD-GFS has demonstrated high internal consistency ($\alpha = .92$) and adequate test-retest reliability ($r = .66$ -. 76 ; Byles et al., 1988; Miller et al., 1985). In the present study, the FAD-GFS demonstrated good-to-excellent internal consistency (pre-FAD-GFS $\alpha = 0.86$; post-FAD-GFS $\alpha = 0.94$; 3-mo. follow-up-FAD-GFS $\alpha = 0.92$).

Objective Measures

Wearable Device – Fitbit Inspire HR (Fitbit Inc, San Francisco, CA, USA). In combination with sleep diaries, the Fitbit Inspire HR was used to assess mean sleep patterns and

sleep-related activity. The Fitbit Inspire HR may be worn consistently for up to 5-days and records wrist movement and heart rate variability to estimate sleep onset, waking, and sleep stages without need for the wearer to actively manipulate the device. Newer, sleep-staging Fitbit devices have been used to supplement subjective sleep data, and have shown good sensitivity values of 0.95-0.96 (Haghighat et al., 2019), but less desirable rates of specificity (0.58-0.69; Haghighat et al., 2019). This device was used to assess broad changes in sleep over time and compare with subjective data to ensure accuracy.

Psychomotor Vigilance Task (PVT). The PVT is a continuous performance task commonly used to objectively assess sleep-related cognitive deficits in vigilant attention (Dorrian, Rogers, & Dinges, 2005). The PVT, administered remotely via Inquisit Web software (Millisecond Software LLC, US) was used in this study, and is analogous to the stand-alone PVT-192, which is not transportable (Reifman et al., 2018). Inter-stimulus intervals are set to vary randomly between 2-10 seconds and subjects are to respond as quickly as possible by clicking the mouse to stimuli flashed on the screen. Higher scores indicate slower response times and greater number of errors. The PVT has demonstrated sensitivity to both total and chronic partial sleep loss at durations of 10 minutes, and has demonstrated high effect sizes for both response speed and number of lapses (lapses are defined as response times >500ms) in individuals who were sleep deprived (Basner & Dinges, 2011). The PVT has demonstrated validity, and reliability for median response times ($ICC = 0.826, p < .0001$) and lapses ($ICC = 0.888, p < .0001$; Dorrian, Rogers, & Dinges, 2005).

Feasibility

Attendance and Adherence. Participant weekly CBT-I-F attendance was tracked and analyzed based upon average percentage of total number of sessions attended. Adherence to intervention

components was tracked via compliance with recommendations for the previous week. After each session, sleep diaries for that week were examined by research assistants and recorded dichotomously as to whether “Yes, the participant was adherent” or “No, the participant was not adherent” (Kamen et al., 2019). Adherence was defined as engaging in weekly recommendations on at least four nights throughout the week (Hebert et al., 2010). An adherence percentage (number of adherent weeks divided by number of sleep diaries returned) was calculated (Kamen et al., 2019).

Acceptability

Acceptability was assessed using four self-report questions derived from similar studies (Christopher et al., 2018); the degree to which the participant liked the program, whether they found the home practice/requirements reasonable, whether they would recommend the program, and whether they would take the program again in the future. Each item is responded to on a 7-point Likert-type scale ranging from 0 to 6, with higher scores indicating greater acceptability.

Fidelity

Fidelity was assessed to ensure therapist adherence to the intervention protocol in each session. An adherence measure of covered session topics, activities, and global instructor skill in presenting material was used for each weekly session. Sessions were audio recorded and two research assistants independently rated each item regarding covered session topics and activities on a 3-point Likert-style scale ranging from 0-2 (0 = *not present*, 1 = *somewhat present*, 2 = *fully present*), and global instructor skill in presenting the material was rated on a 3-point Likert-style scale ranging from 0-2 (0 = *none*, 1 = *minimal*, 2 = *adequate*).

Procedures

General Overview

All study procedures were approved by the Pacific University IRB. Recruitment for this study began in March 2021. Several chiefs of moderate-sized, urban fire departments within the Pacific Northwestern U.S. were contacted to determine potential interest in the study and to gain approval on dissemination of study information within their department. Following approval from administration, information regarding the study was disseminated through in-station fliers, email, and snowball recruitment. Potential participants were encouraged to email or call the primary study investigator regarding interest and to arrange a screening appointment.

During the screening appointment, participants were asked to provide written informed consent and were subsequently screened for eligibility. If eligible, participants were instructed to complete baseline assessments via computer (PSQI, DASS-21, PCL-5, SF12v2, FAD-GFS, and PVT). Following completion of these baseline assessments, participants were randomly assigned (via 1.5:1 ratio (CBT-I-F:WLC) to one of two groups (CBT-I-F or WLC). The sample was intended to be stratified by gender; however, only male participants enrolled in this study, thus stratification was not needed. Participants assigned to the WLC group were provided the opportunity to participate in CBT-I-F following the study. Fitbit Inspire HR devices and the CSD (sleep diaries) were distributed to participants randomized to the CBT-I-F group. Participants were instructed to wear the Fitbit throughout the duration of the 6-week intervention and the following two weeks to collect objective data on enduring sleep changes. Participants randomized to the CBT-I-F group were asked to complete the CSD for the week prior to the first CBT-I-F session and continue to complete it throughout the intervention. The group CBT-I-F session outline is presented in Table 2. Sessions were led by two graduate student therapists

under the supervision of a licensed clinical psychologist experienced in CBT-I interventions and studies involving first responders.

As part of the aim to increase the feasibility of this intervention for firefighters, two CBT-I-F group sessions were offered concurrently. The two group sessions were provided each week on back-to-back days (i.e., Monday and Tuesday) and were designed to cover the same material. Participants were instructed to attend only one of the two sessions each week, but were allowed the discretion to choose which they would attend. This addition to the study design was included due to the nature of firefighter shift schedules (i.e., 24-hour rotating shifts), that might otherwise limit ability to attend.

Participants in both groups (CBT-I-F and WLC) completed a similar battery of assessments (PSQI, DASS-21, PCL-5, SF12v2, FAD-GFS, and PVT) following the last CBT-I-F intervention session. Measures were repeated for both groups at 3-month follow-up, with the exception of Fitbit measurements.

Setting

Due to the COVID-19 pandemic, the setting for this study was limited to an online format. Both assessment and group therapy sessions were held via data-protected Zoom and disclosures regarding confidentiality of participation via web-based platforms were provided during the informed consent. Several meta-analyses (Peter et al., 2019; Ritterband et al., 2017; Ye et al., 2016) have demonstrated that online dissemination of CBT-I is as efficacious as in-person.

Cognitive Behavioral Therapy for Insomnia for Firefighters (CBT-I-F)

CBT-I-F is a six-session weekly group protocol. Details are presented in Table 2. The content of each session was developed from the combination of protocols described in Manber and Carney (2015), Perlis et al. (2008), and Taylor et al. (2019). Themes and occupational adaptations provided in this intervention included the following.

Stimulus Control. Stimulus control requires strict adherence to behavioral conditioning techniques to re-teach the body and brain to respond to the bed as a cue for sleep (Manber & Carney, 2015). Stimulus control was adapted to account for on-duty versus off-duty nights. While on-duty, firefighters are unable control the time in which they consistently wake every morning due to the unpredictability of emergency calls. They may also experience nights where they are required to be awake for extended periods due to these demands. For these reasons, having a consistent wake time was emphasized for mornings that they are off-duty, rather than mornings for which they are at work, and napping was not expected to be eliminated entirely, rather it was encouraged only for short durations (i.e., <30 minutes; NSF, 2020e), early in the day following sleepless shifts.

Sleep Restriction. Sleep restriction intentionally restricts sleep to the hours in which one believes that they can sleep soundly through the night. For some individuals, sleep restriction significantly reduces the amount of time they are permitted to rest in bed (Manber & Carney, 2015). While sleep restriction is often a main component of CBT-I, several protocols recommend it as an optional component for high-risk populations as it can increase daytime sleepiness and heighten the risk for injury in populations needing to operate machinery and sustain high levels of alertness (Manber et al., 2014; Murphy et al., 2015; Spielman et al., 2011). Although these risks generally only occur during acute periods following the initiation of sleep restriction (due to initial increases daytime sleepiness), it may be too risky for certain populations even short term.

Due to the high-risk nature of firefighting, sleep restriction was discussed in session, but not required unless firefighters were able to take long vacation periods during which daytime sleepiness would not impact job safety.

Counter-Arousal Skills and Cognitive Restructuring. Counter-arousal skills and cognitive restructuring aim to reduce the impact of cognitive arousal before bed and during periods of awakening throughout the night (Manber & Carney, 2015). Examples of counter-arousal skills that were used in this intervention included scheduled worry time, constructive journaling, diaphragmatic breathing, and body scan meditation (Manber & Carney, 2015; Perlis et al., 2008). Because firefighters must respond to emergency calls at all hours, arousal is unavoidable and common during evening hours and throughout the night (Ferguson et al., 2016; Kim et al., 2019; Paterson et al., 2016). Counter-arousal skills and cognitive restructuring were emphasized given the prevalence of hyperarousal in firefighter populations, and these techniques were tailored to firefighter concerns, such as trying to return to sleep following emotionally arousing calls, mitigating anxiety about functional impairment the subsequent day in relation to work or at-home duties, and establishing calming activities conducive to sleep at both the fire station and at home.

Sleep Hygiene. Sleep hygiene skills are recommendations that can be carried out at various times throughout the day to support sleep quality at night (Manber & Carney, 2015). Examples of relevant sleep hygiene skills that were included in the intervention were daily exercise, cutting down on caffeine products, avoidance of alcohol in the evenings, avoidance of heavy meals before bed, and creating a quiet and dark sleeping environment (Perlis et al., 2008). Firefighters face rotational changes and 24-hour shifts (Cohen & Plecas, 2013). This may make some sleep hygiene skills difficult to implement as it is encouraged to establish daily routines

and have the same bedroom environment conditions night-to-night. While traditional sleep hygiene skills were recommended, problem solving for occupational-related factors were a key component of group discussion.

Additional Adaptations. Due to the aim to increase feasibility and acceptability of CBT-I within a firefighter population, additional adaptations to the protocol included culturally normative language and discussions around topics that pertain specifically to the career of firefighting. Christopher et al. (2016, 2018) found promising benefits of adapting traditional therapeutic interventions to first responder populations, including demonstrating high acceptability and adherence rates. Adaptations for this study included, but were not limited to: discussion of how sleep affects the ability to be responsive and alert while responding to emergency calls; how skills may be implemented successfully across two separate sleeping environments (home and fire station); creating a routine for returning to bed after inevitable nighttime sleep interruptions; and additional implications of how traumatic experiences and family-conflict can influence sleep and ability to adhere to recommendations.

Actigraphy

Fitbit Inspire HR wrist-worn devices were loaned to eligible participants assigned to the CBT-I-F group during the screening appointment. Instructions were provided, and participants were asked to wear the Fitbit for the duration of the intervention plus two weeks following the intervention (i.e., worn for a total of 8-weeks). Fitbits and Fitbit accounts were set up using generic emails using only participant ID numbers. All Fitbit-related data were deidentified. Data was synced weekly to the Fitbit accounts and analyzed by the study investigator(s) to inform group CBT-I-F treatment and assess sleep patterns for each participant.

Data Analysis

Data storage and analyses were conducted using IBM SPSS Statistics Software 28 (SPSS 28.0; IBM Corp., 2021). Data was cleaned prior to analysis to assess for missingness and outliers. Missingness was analyzed via a Missing Values Analysis. Little's MCAR test was used to reveal non-significant p -values, indicating that data are likely missing completely at random. Missing Values Analysis using a Little's MCAR test revealed non-significance across self-report measures ($p = .981$), PVT data ($p = .309$), and Fitbit data ($p = 1.00$) indicating that data are likely missing completely at random. An advantage of linear mixed effect modeling is that it does not require complete data over occasions (Tabachnick & Fidell, 2019), thus missing data on continuous repeated outcome measures (i.e., self-report data) were handled via maximum likelihood estimation. Maximum likelihood estimation, however, requires data to be missing at random, which necessitated the inclusion of analyses via Little's MCAR test (Li, 2013). Only observed data were used to analyze attendance and attrition rates. Missing data of >5% appearing on measures of acceptability was proposed to be handled via multiple imputation methods (Jakobsen et al., 2017; Tabachnick & Fidell, 2019), as missing data on acceptability measures was not otherwise accounted for via maximum likelihood estimation due to its administration at only one time point. However, upon visual inspection, there was only one participant who had missing data for acceptability measures, and this participant was also non-adherent to much of the protocol. SPSS Version 28.0.0 does not allow multiple imputation to be run for data with less than 10% missingness, as such, analyses for acceptability measures were completed with only available data as an alteration to the data analysis plan. Missing data of >5% appearing on Fitbit data was handled via multiple imputation methods and analyses were run with imputed data and without per best practice guidelines (Jakobsen et al., 2017; Tabachnick & Fidell, 2019).

Univariate outliers were assessed on each subscale and indicated by standardized values (z scores) that are outside the absolute value of 3.29. Two univariate outliers were identified, one on the PVT Baseline Mean Response Time item, and one on the PVT Number of False Starts item. Because of the small sample size, these values were determined to be critical to the analysis and as such, were transformed via winsorizing (Ghosh & Vogt, 2012). After winsorizing univariate outliers, multivariate outliers were assessed for via Mahalanobis distance and resulting outliers were screened for correlation with other measures (Tabachnick & Fidell, 2019). If true multivariate outliers remained after transformation, they were planned to be deleted from the analysis, as is best practice reported by Tabachnick and Fidell (2019). The data produced no multivariate outliers. To investigate normality, a Shapiro–Wilk test was performed provided it has more power to detect nonnormality in smaller sample sizes (Mishra et al., 2019); transformations are to occur according to significance at the $p < .05$ level and provided they do not produce skewness or kurtosis at different time points in the dataset (Field, 2018). The Sharpiro-Wilk test demonstrated significance values of less than .05 for the following variables: DASS-21 Anxiety, FAD-GFS, PCL-5, PVT Mean Response Time, PVT Average Lapse Time, Fitbit Mins Asleep, Fitbit Mins Awake, Fitbit TIB, Fitbit Mins Rem, Fitbit Mins Light, and Fitbit Mins Deep. The variables of DASS-21 Anxiety, FAD-GFS, PCL-5, and PVT Average Lapse Time were improved when transformed via a Square Root transformation. PVT Mean Response Time, Fitbit Mins Asleep, Fitbit Mins Awake, Fitbit TIB, Fitbit Mins Rem, Fitbit Mins Light, and Fitbit Mins Deep did not improve with transformation, thus were left un-transformed for data analysis and analyses using these variables were treated with caution.

Data was assessed for the assumptions of linearity and homogeneity of variance. Linearity was assessed by examining graphical plots of the model residuals versus the predictors

(Tabachnick & Fidell, 2019). While multilevel models (MLMs) do not require assumption of homogeneity of variance, it was assessed due to the simplistic nature of this specific MLM design by conducting a Levene's Test (Field, 2018). If p values are greater than 0.05 then the assumption of homogeneity has been met, but are to be interpreted cautiously as recommended in Field (2018). Levene's Test revealed only one failing variable, PCL-5 3-month follow-up (SQRT). Cronbach's alpha internal consistency was run on all study measures.

To analyze Hypothesis 1, descriptive statistical analyses were conducted at a post-intervention, including completion percentages (%), means (M), and standard deviations (SD). As functions of feasibility, attendance and attrition rates were used, and based upon evidence from similar studies of first responder interventions (Christopher et al., 2018) and CBT-I interventions (Koffel et al., 2015), an attendance rate of $\geq 80\%$ and attrition rate of $\leq 20\%$ (Hebert et al., 2010; Ong et al., 2008) were predicted. Adherence to recommended home practice was assessed based on a completion percentage of returned sleep diaries [CSD], and based upon evidence from similar studies using CBT-I interventions, a completion rate of $\geq 60\%$ was expected (Kamen et al., 2019; Matthews et al., 2013). Fidelity was assessed according to the expectation that the therapist would adhere to the intervention protocol in both sessions provided each week. Two-way random-effects models evaluating intraclass correlation (ICC) were used to demonstrate inter-rater consistency for mean ratings of covered session topics, covered activities, and global instructor skill in presenting material (Christopher et al., 2018; Koo et al., 2016). ICC coefficients close to one indicate greater agreement among raters (Koo et al., 2016). The following qualitative interpretations of ICC values were used: $ICC < 0.50$ indicates poor reliability, $ICC = 0.5 - 0.75$ indicates moderate reliability, $ICC = 0.75 - 0.9$ indicates good reliability, and $ICC > 0.9$ indicates excellent reliability (Koo & Li, 2016). Means (M) and

standard deviations (*SD*) are reported with higher mean ratings indicating greater coverage of session material and therapist skill.

To analyze Hypothesis 2, descriptive statistical analyses were conducted at post-intervention, including means (*M*), and standard deviations (*SD*). Based upon evidence from similar studies investigating the acceptability of interventions within first responder populations (Christopher et al., 2018), a mean acceptability rating of ≥ 4 (on a scale of 0-6), was predicted. Ratings of ≥ 4 (on a scale of 0-6) indicate that the intervention was generally accepted.

To analyze Hypothesis 3, change in primary and secondary preliminary outcomes across the three time points (baseline, post-intervention, 3-month follow-up), linear mixed effect models, using maximum likelihood estimation analyses were conducted to test group, time, and interaction effects for continuous dependent variables based on a 2 (intervention groups) x 3 (assessment time points) design. Data were analyzed using an intent-to-treat model. The linear mixed effect model design includes participant as a random effect, and group, time, and group-by-time interaction as fixed effects. Determination whether to include random intercepts or slopes within the model were made. The overall fit of the model was tested using a chi-square likelihood ratio test (Field, 2018). Repeated measures were fitted to a covariance structure to reduce Type I and II errors, and then the model was run with each suspected covariance structure, using appropriate goodness-of-fit indices (Akaike's information criterion (AIC); Hurvich and Tsai's criterion (AICC); Bozdogan's criterion (CAIC); Schwarz's Bayesian criterion (BIC; Field, 2018). Analyses were conducted separately to investigate each continuous dependent variable: sleep quality-subjective [PSQI; primary outcome; separating work and home], depression [DASS-21 Depression; secondary outcome], anxiety [DASS-21 Anxiety; secondary outcome], stress [DASS-21 Stress; secondary outcome], PTSD [PCL-5; secondary

outcome], general mental health [SF12v2 – Mental Component Summary; secondary outcome], general physical health [SF12v2 – Physical Component Summary; secondary outcome], vigilant attention [PVT; secondary outcome], and family functioning [FAD-GFS; secondary outcome] relative to participants in the waitlist control group. Results are reported on means (see Table 3), standard deviations, *p*-values, *F*-values, effect sizes, and 95% confidence intervals. Post-hoc analyses were conducted by using non-parametric tests; the Friedman Test to investigate within-group changes over the three timepoints, and the Mann-Whitney U Test to investigate between-group differences at the three different timepoints. Sleep quality-objective [Fitbit Inspire HR; primary outcome] was analyzed by deriving the mean values each week for the variables of minutes asleep, minutes awake, number of awakenings, total time in bed, minutes in REM sleep, minutes in light sleep, and minutes in deep sleep, then conducting visual inspections of plotted points and repeated measures ANOVAs to analyze relative changes over time. All assumptions were tested prior to analyses, including: 1) the dependent variable should be interval or ratio data; 2) the participants are randomly sampled from the population; 3) the dependent variable is normally distributed in the population at each level of the independent variable; and 4) the variability in the differences between any pairs of groups should be the same as differences between any other pairs of groups. Data on sleep quality assessed via the Consensus Sleep Diary was analyzed by deriving the mean values each week for each participant on the variables of sleep onset latency, number of awakenings, total sleep time, sleep quality, refreshment by sleep, number of naps, alcoholic beverages, and caffeinated beverages. Mean values were graphed for visual inspection.

The linear mixed effect model was chosen for data analyses method based on several conditions: 1) similar study designs (Christopher et al., 2018; Norell-Clarke et al., 2015; Taylor

et al., 2014) have considered linear mixed effect modeling more appropriate than analysis of variance analyses (ANOVAs) for longitudinal data because there is often correlation between repeated measures at the individual level and ignoring this phenomenon may increase the risk of Type I errors (Norell-Clarke et al., 2015), 2) it is not necessary to have complete data or equal numbers of cases when applying multilevel-linear modeling (MLM; Tabachnick & Fidell, 2019), 3) sphericity need not be assumed (Field, 2018), and 4) Boisgontier and Cheval (2016) suggest that research methodology is moving toward the implementation of mixed modeling rather than ANOVAs, due to the aforementioned advantages, as well its ability to treat subjects/conditions as random effects allowing results to generalize.