Published in final edited form as:
Addict Res Theory. 2014 ; 22(5): 407–415. doi:10.3109/16066359.2013.868446.

Research staff training in a multisite randomized clinical trial: Methods and recommendations from the Stimulant Reduction Intervention using Dosed Exercise (STRIDE) trial

Robrina Walker, PhD[1], **David W Morris, PhD**[1], **Tracy L Greer, PhD**[1], and **Madhukar H Trivedi, M.D.**[1]

[1]Department of Psychiatry, University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, Texas, USA 75390-9119

Abstract

Background—Descriptions of and recommendations for meeting the challenges of training research staff for multisite studies are limited despite the recognized importance of training on trial outcomes. The STRIDE (STimulant Reduction Intervention using Dosed Exercise) study is a multisite randomized clinical trial that was conducted at nine addiction treatment programs across the United States within the National Drug Abuse Treatment Clinical Trials Network (CTN) and evaluated the addition of exercise to addiction treatment as usual (TAU), compared to health education added to TAU, for individuals with stimulant abuse or dependence. Research staff administered a variety of measures that required a range of interviewing, technical, and clinical skills.

Purpose—In order to address the absence of information on how research staff are trained for multisite clinical studies, the current manuscript describes the conceptual process of training and certifying research assistants for STRIDE.

Methods—Training was conducted using a three-stage process to allow staff sufficient time for distributive learning, practice, and calibration leading up to implementation of this complex study.

Results—Training was successfully implemented with staff across nine sites. Staff demonstrated evidence of study and procedural knowledge via quizzes and skill demonstration on six measures requiring certification. Overall, while the majority of staff had little to no experience in the six

Corresponding Author: Robrina Walker, PhD, Department of Psychiatry, University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd., Dallas, TX 75390-9119; Phone: 214.645.6980; Robrina.Walker@utsouthwestern.edu.

DECLARATION OF CONFLICTING INTERESTS

Drs. Walker and Morris have no conflicts to disclose. Dr. Greer has received research funding from NARSAD and is a paid consultant for H Lundbeck A/S. Dr. Trivedi has received research support from the Agency for Healthcare Research and Quality (AHRQ), Corcept Therapeutics, Inc., Cyberonics, Inc., Merck, National Alliance for Research in Schizophrenia and Depression, National Institute of Mental Health, National Institute on Drug Abuse, Naurex, Novartis, Pharmacia & Upjohn, Predix Pharmaceuticals (Epix), Solvay Pharmaceuticals, Inc., Targacept, and Valient, and consulting and speaker fees from Abbott Laboratories, Inc., Abdi Ibrahim, Akzo (Organon Pharmaceuticals Inc.), Alkermes, AstraZeneca, Bristol-Myers Squibb Company, Cephalon, Inc., Evotec, Fabre Kramer Pharmaceuticals, Inc., Forest Pharmaceuticals, GlaxoSmithKline, Janssen Pharmaceutica Products, Libby, LP, Johnson & Johnson PRD, Eli Lilly & Company, Meade Johnson, Medtronic, Neuronetics, Otsuka Pharmaceuticals, Parke-Davis Pharmaceuticals, Inc., Pfizer Inc., Sepracor, SHIRE Development, Sierra, Tal Medical/Puretech, Transcept, VantagePoint, and Wyeth-Ayerst Laboratories.

measures, all research assistants demonstrated ability to correctly and reliably administer the measures throughout the study.

Conclusions—Practical recommendations are provided for training research staff and are particularly applicable to the challenges encountered with large, multisite trials.

Keywords

training recommendations; research staff training; multisite trials; training; certification

INTRODUCTION

Research assistants for clinical research trials often must possess a broad range of interpersonal and technical skills in order to adeptly administer the variety of measures required to evaluate primary and secondary outcomes. For example, research assistants (RAs) often are expected to administer simple demographic questionnaires as well as semi-structured diagnostic interviews requiring more advanced clinical knowledge and interviewing skills. The impact of training methodology on clinical trial outcomes, especially diagnostic and other outcomes requiring judgment, has been examined in regard to its efficacy (Mulsant et al., 2002; Targum, 2006) and potential impact on failed trials (Kobak, Engelhardt, Williams, & Lipsitz, 2004). Despite the clear importance of training, descriptions of training provided in clinical trials are usually not adequately reported (Mulsant et al., 2002) and are typically limited to one to three paragraphs. Articles devoted to training focus solely on inter-rater reliability aspects of single measures (e.g., Jeglic et al., 2007; Kobak, Engelhardt, & Lipsitz, 2006; Kobak, Lipsitz, Williams, Engelhardt, & Bellew, 2005; Rosen et al., 2008), to the exclusion of the training provided on study procedures not specific to a given measure (e.g., using data capture systems, recruitment methods). Furthermore, effectively training research staff for multisite randomized clinical trials poses additional challenges due to geographic distribution of staff, difficulty scheduling training session thus leading to extended training timelines, hiring decisions resulting in staff with varying skills, and limited opportunities for trainers to work face-to-face with staff when needed. However, contemporary recommendations for meeting training challenges in multisite studies are not present in the literature despite the recognized importance of training on trial outcomes.

The current paper conceptually describes the three-stage training and certification process used to train RAs for the STRIDE (Stimulant Reduction Intervention using Dosed Exercise) trial (Trivedi et al., 2011), a multisite randomized clinical trial conducted at nine community addiction treatment programs within the National Drug Abuse Treatment Clinical Trials Network (CTN). STRIDE was conducted by the Lead Investigator’s Team based in Dallas, TX in collaboration with the NIDA-sponsored clinical coordinating center (CCC) and data and statistics center (DSC2) based in Bethesda, MD, collectively known as the Lead Team (LT). The nine study sites were spread across three U.S. time zones. The trainers (RW and DWM, with additional collaboration with the CCC) developed the research staff training and certification process presented here. Conclusions and recommendations for training and evaluating research staff in large multisite studies are provided.

METHOD

STRIDE evaluated the impact of exercise versus health education as augmentation treatment strategies to usual care on drug use outcomes in individuals with stimulant use disorders. Participants ($N = 302$) began residential substance use treatment at the participating study site, provided informed consent, and were randomized to a treatment arm. Research and intervention visits occurred three times per week for the first three months and once weekly for the final six months. The recommended staffing for each site required the two RAs to be certified as a back-up interventionist for one of the treatment arms to ensure each role had back-up coverage at any given time, adding to the training burden of the RAs. Additional information on study design is provided elsewhere (Trivedi et al., 2011).

Three distinct training periods were used in STRIDE. First, pre-training was conducted remotely using various methods over a two- to four-week period. In-person training then occurred during a three-day training meeting. Finally, post-training was conducted remotely over a two- to four-week period. During this time RAs finalized local standardized operating procedures specific to their sites' needs and were certified to administer measures. The trainings first focused on non-protocol specific, basic data collection procedures, then on protocol-specific mid-level skills, and later focused on complex protocol-specific topics, such as assessment-specific training. This graduated training process optimized the limited in-person training time by focusing on complex topics and incorporating experiential learning. Training was conducted with the first wave of four sites and seven months later with the second wave of five sites. Prior to training, staff experience was evaluated via an emailed form that allowed trainers (RW, DWM) to understand raters' experience, with adequate experience being defined as having a minimum of two years' experience administering a given measure once per month.

Pre-training Period

The pre-training period was designed to ensure all staff attended the training meeting with similar knowledge in basic research methodology (e.g., informed consent process, regulatory requirements and documentation) and selected aspects of the study protocol. Protocol-specific pre-training included, for example, recruitment and retention procedures, the medical screening visit, and administration of select measures. Pre-training sessions were conducted using phone calls in which emailed materials were used during the calls, webinars in which polling questions in which attendees logged their responses were used to gauge real-time learning, and self-paced reading of manuals. During this period staff also developed site specific recruitment, enrollment, and retention procedures and practiced administering measures with colleagues.

In-Person Training Meeting

The three-day in-person training meeting was designed to provide in-depth training on complex protocol-specific tasks. The first two days addressed research and intervention procedures for all team members while the third day was for research assistant procedures. A combination of didactics, live demonstrations, role plays, and experiential learning was used throughout the training meeting. Trainers used didactics with supporting PowerPoint slides showing case report forms (CRFs) and other study documents. Live demonstration of the electronic data system's general navigation and study-specific functionality was provided. Training on interviewer-administered assessments was conducted by expert demonstration and role play with trainees coding along. Use of supplies the sites had not yet received (e.g., urine drug screens, pregnancy tests) were demonstrated. Exercise interventionists and their back-ups previously had attended a two day training focused on exercise procedures and dosage; additional STRIDE-specific training for all interventionists occurred during breakout sessions, including a role play of key components of the first intervention session.

At the end of each 30 to 60 minute training session throughout the three days, questions assessing key knowledge areas were displayed on the projector screen and trainees used personal response devices to input their individual answers which were then (1) recorded by the system for later review and (2) displayed collectively on the projector. The latter provided instant feedback on learning and indicated areas needing clarification. Trainees who answered the polling questions accurately and the fastest received a reward representative of items study participants would receive or use during the study (e.g., note pad and pen, heart rate monitor) and were provided to help increase staff engagement during the training meeting.

Post-training Period

The post-training period was designed to allow time for the following: evaluation and certification of staff to administer six pre-defined assessments, final preparations by sites for study implementation, and standardized patient visits to the sites. The Lead Team identified post-training needs by reviewing incorrectly answered polling questions and by compiling questions posed by site staff. For staff that could not attend training sessions, make up sessions were provided via telephone during the post-training period.

Post-training typically was informal and occurred via email and telephone, with responses distributed to the national study team via conference calls and emails. Two formal trainings were held via webinar. The first webinar trained RAs on software used to administer the substance use modules of the World Health Organization Composite International Diagnostic Interview (CIDI) version 2.1 (1997), a standardized diagnostic measure RAs were certified on by a CTN CIDI trainer affiliated with sites' corresponding Regional Research and Training Center. The second webinar explained the CTN's process of site endorsement and quality assurance monitoring throughout the study. A sample administration video for the MINI International Neuropsychiatric Interview (MINI; Sheehan et al., 1997) was provided for site staff after having been trained on the MINI's components at the training meeting. Finally, a separate CTN team conducted standardized patient walkthroughs that allowed staff the opportunity to assess preparedness and practice study procedures (Fussell, Kunkel, McCarty, & Lewy, 2011). The standardized patient team provided constructive feedback to RAs at the end of their visit and a summary report to the LT.

Competency evaluations

Staff competency was evaluated with quizzes that objectively measured knowledge pertaining to STRIDE's objectives and procedures and the LT trainers formally evaluated staff skills via practica and audio recordings. Competency was evaluated throughout the three training stages, culminating in certification to perform the RA role.

Quizzes—Polling questions delivered via the personal response devices during the training meeting served as the STRIDE study quiz. A few additional quizzes were administered via email during the pre- and post-training meeting periods. A minimum score of 80% was required for all STRIDE site staff. However, trainees scoring < 100% were emailed requests to answer the incorrect items by two weeks after the training meeting to help ensure full understanding of the study and procedures.

Skill demonstration

Data system: The DSC2 developed a practicum to objectively evaluate RA's completion of key tasks within the system (e.g., form creation, missing item requests) which also ensured staff practiced navigating key areas within the EDC system. Written scenarios likely to be encountered during the study were provided and staff determined how and where to enter the data into the data system. An emailed certificate was provided when staff demonstrated proficiency.

Certification on interviewer-administered measures: Certification was required for six study measures, each of which were selected due to their importance and administration difficulty, as the measures required a variety of clinical interviewing skills, attention to detail, multi-tasking, and a psychiatric diagnostic knowledge base [see Table 1; a complete measures list is in Trivedi et al. (2011)]. The primary outcome measure, the Timeline Followback (Baca-García et al., 2001) is an interviewer-guided retrospective assessment of substance use commonly used in CTN trials but slightly modified to capture additional data for STRIDE. Two standardized neurocognitive measures, the Stroop Color and Word Test (Stroop; Golden, 1978) and the Wechsler Test of Adult Reading (WTAR; Wechsler, 2001) that requires scoring the pronunciation of up to 50 progressively more difficult words, assessed impulsivity and premorbid intelligence, respectively. Finally, three measures were semi-structured interviews requiring interviewer knowledge and judgment for depression symptoms [Quick Inventory for Depressive Symptomatology, Clinician-Rated version (QIDS-C_{16}); (Rush et al., 2003)] stimulant withdrawal [Stimulant Selective Severity Assessment (SSSA) based on the Cocaine Selective Severity Assessment (Kobak, 2010)] and Axis I psychiatric diagnoses (MINI; Sheehan et al., 1997). Of note, even after RAs were certified to administer the MINI, it was administered under the supervision of a clinician at each site with the exception of one back-up RA who had extensive clinical experience.

Achieving competency on the QIDS-C_{16} required each RA to code two “gold standard” recorded administrations, submit ratings for evaluation, and to score within two points of the gold standard total score and within one point on each individual item on both QIDS-C_{16} certification videos. Achieving competency on the remaining measures required RAs to audio record themselves administering measures to a “mock participant,” most often a colleague, clinic staff member, or even volunteer clients, who role played a stimulant user. Recordings were sequential to allow staff to incorporate constructive feedback during the second recording. Competency recordings were evaluated by the content expert trainers, each of which was responsible for certifying staff on three of the six measures throughout the trial to ensure consistent reviews. The trainers listed to the recorded administration, scored the measures, and compared their expert scores to the RAs' submitted scores. Of note, trainers also evaluated administration skills, such as interviewing skills and documentation. Achieving competency on the TLFB required RAs to elicit all substance use episodes and code all items correctly. For the Stroop, RAs were required to score within one point of the expert's scores for the three scales and achieve perfect scores on one recording. For the MINI, RAs were expected to achieve perfect diagnostic accuracy. For the WTAR, RAs were required to score within two points of the expert's total score. Finally, for the SSSA, RAs were required to score within two points of the expert's total score and within one point on each individual item.

All feedback was provided via email and local site supervisors were copied to ensure they were aware of staff progress. Additional feedback was provided via telephone when necessary to ensure full understanding. For RAs who could not achieve certification, trainers provided remediation training, made additional individualized recommendations for further training and practice depending, and required submission of additional audio recordings,. Examples of recommendations made included more practice with provided scenarios, working with a colleague who had been certified, working with the STRIDE trainer via telephone, and identifying a local supervisor who could assist in person. Site supervisors were copied on these emailed communications to ensure recommendations were followed. If significant deficiencies continued, the STRIDE trainer contacted the staff member's supervisor directly to discuss training or staffing options.

Re-certification on interviewer-administered measures: A similar process was followed to re-certify RAs on the six measures. Approximately every 12 months, RAs were asked to record administrations of all six measures, with the participant's permission. Remediation efforts undertaken to re-calibrate with the trainers were similar to those during the original certification process. RAs who demonstrated significant deficiencies on the first recording or could not demonstrate proficiency with a second recording were de-certified until the deficiencies were corrected.

Ongoing training throughout the trial

Informal, ongoing training was provided as needed throughout the trial, and training needs were identified by several methods. First, routine conference calls, emails, and individual phone calls revealed areas requiring more training. Second, reports from national quality assurance monitors from the CCC and local quality assurance monitors from the Regional Research and Training Center affiliated with each site identified procedures staff were not implementing consistently. Training needs also were identified by the national DSC2 data managers who fielded data questions from the sites, identified problems with data entry by reviewing queries built into the EDC system, and identified unique data entry problems through systematic data reviews. The DSC2's daily trial progress report with standardized metrics used across all CTN studies (e.g., screening rate, reasons for exclusion at screening) and metrics requested by the LT (e.g., percent of primary outcome data collected to date) revealed areas needing more attention and retraining. Finally, reports and logs of safety events and protocol departures also identified training needs. Ongoing training was provided via several forums. Answers to questions and further informal training were provided directly to sites via teleconferences (i.e., national conference calls, calls with site teams, calls with individual staff members at sites, calls with site supervisors), emails, and in-person site visits.

RESULTS

Quiz Results

The majority of staff passed the main STRIDE quiz at the minimum required 80% rate. The majority of staff who scored less than 100% answered items correctly during the second opportunity. Trainees who answered incorrectly the second time were emailed an explanation and rationale for the correct answer. Local site supervisors were included on these emails and, of note, supervisors were separately alerted when staff answered a significant number of items incorrectly on the quiz or items specific to their study role so local remediation training on deficient areas could be facilitated. Thus, although the quizzes were administered centrally by the LT's trainers, local site supervisors also assisted with remediation efforts. Ultimately, all staff members passed the quizzes required for their given role.

Skill demonstration

Data system: All staff passed the EDC evaluation with relative ease. However, there were several tasks related to electronic form creation and data entry for the primary outcome measure that was difficult for some to implement. In those instances, RAs either contacted the DSC2's data manager for clarification and assistance or waited to receive feedback on the materials submitted for review and scoring. Staff received feedback and guidance on tasks completed incorrectly and all were asked to continue practicing within the training platform of the EDC's system until recruitment began.

Certification on interviewer-administered measures: The prior experience survey RAs completed during pre-training indicated the majority had minimal to no prior experience with STRIDE measures. Thus, trainers reviewed the certification audio recordings for evidence of scoring to the gold standard and correct administration, as well as interviewing style and correct documentation. As such, scoring accuracy as well as the overall gestalt of RAs' administration were considered during the certification process. A summary of areas assessed, problems identified, and corrective actions recommended to achieve certification are provided in Table 2.

Scoring: Primary concerns regarding scoring on QIDS-C_{16} centered on systematic under or over rating. Additional QIDS-C_{16} "gold standard" videos were provided for RAs (<25%) unable to pass one or more of the first set of videos. For the TLFB, the majority of RAs elicited and documented substance use data with perfect accuracy, with a minority of RAs needing additional training on formulating questions to gather accurate data. These problems were generally resolved in the second recording. For the Stroop, RAs generally achieved excellent scoring within the first two recordings, with a minority of RAs (<15%) submitting a third recording for certification. For the MINI, experienced RAs with prior mental health training performed the best. Difficulties on the MINI primarily stemmed from a lack of diagnostic experience, which led to RAs not eliciting additional information via impromptu and thorough follow-up questions.

The advanced words on the WTAR were more problematic for RAs to score properly and a third recording was required for less than 25% of the RAs. Online dictionary websites with audio pronunciations were provided to assist with correct scoring of particularly difficult words.

The SSSA, which was developed as a clinician-rated scale and does not include suggested interview questions or prompts on the form, required significant remediation (> 50% of RAs) for errors in scoring. The greatest source of error was using the Likert scale anchors as structured interview prompts, resulting in a higher likelihood of participants selecting those anchors, rather than asking appropriate open-ended questions. If left uncorrected, this would have reduced variability of the scale.

Administration: RAs evidenced the most difficulty mastering the nuances of administering the semi-structured interviews. Staff could phrase questions in several ways to gather the necessary information; however, poorly phrased questions would not elicit the necessary information and unnecessarily lengthen interviews. Examples of poor phrasing for the TLFB included incorrectly asking for quantity of substances used and not using suggested techniques. For the MINI, administration problems were generally in the form of lack of thorough follow-up questioning that gathered sufficient information to make a clinical judgment in the final rating. As the trainers encountered these problems, they provided specific feedback detailing the RA’s statement in the recording and suggesting alternative ways to ask the question to improve the administration method on the next recording.

The Stroop and WTAR required staff to follow fully standardized administration instructions. The WTAR’s directions were printed directly on the form and are brief; few staff demonstrated difficulty with correctly administering the WTAR. In contrast, the Stroop’s directions are in the user’s manual. The majority of staff needed additional training and coaching on administration principles of psychological testing. For example, many required coaching on having the verbatim directions available to read, rationale for stating the directions without deviation, speaking rate and the need for pauses when giving directions, timing the task properly, and managing the multiple administration demands of the task.

Interviewing skills: Trainers also provided feedback on issues observed related to qualitative interviewing skills and style. For example, for RAs who administered measures correctly but quickly, trainers provided feedback to slow down their rate of speech or to allow more time for participants to respond after asking a question. Since the SSSA does not contain interview prompts or comprehensive scoring guidelines, RAs were provided feedback on the need and rationale for formulating open-ended questions rather than reading the verbatim Likert scale response option anchors. Feedback on interviewing style also included suggestions for transitioning between measures, reminders to encourage and thank participants, and other “customer service” types of skills that impact rapport with participants, and, ultimately, study retention and data quality.

Documentation: There were several instances in which the trainers gave feedback about documentation and proper error correction procedures, and these reviews were most advantageous on the TLFB and the SSSA. The TLFB’s CRF included 8 columns and 19 rows, with each intersection of them containing 1 to 4 data elements. Staff were initially trained on, and manuals also contained information on, documentation shortcuts deliberately developed to streamline form completion, for example, if a participant reported not using substances on a given day. However, some staff needed additional coaching on these shortcuts, which was only evident during review of their certification submissions. Documentation review was also helpful for the SSSA’s visual analog scale in which staff had to interpret and record the numerical equivalent of the mark on the scale.

Re-certification on interviewer-administered measures: Most RAs earned re-certification at approximately 12 month intervals by submitting one audio recording for each measure. A few RAs who had difficulty demonstrating proficiency and obtaining certification at the start of the trial (i.e., more than two recordings were required to be certified for a given measure) also demonstrated some difficulty obtaining re-certification.

In consultation with sites’ local supervisors, trainers de-certified one RA who had difficulty earning re-certification rather than devoting time to remediation training because the local supervisor determined adequate staffing was in place to cover the study tasks. A few RAs’ (<10%) re-certification recordings demonstrated adequate skill to continue administering the measures to study participants, but they needed re-calibration on asking adequate follow-up questions or not deviating from structured interview questions. Staff members were provided constructive feedback and were conditionally re-certified with a requirement to submit one to two additional recordings to demonstrate proficient skills for full re-certification.

DISCUSSION

This paper conceptually described the methods utilized to train RAs in the STRIDE trial. Training was conducted using a three-stage process—pre-training, training meeting, and post-training— employed deliberately to allow staff sufficient time for distributive learning, practice, and certification leading up to study implementation given the complexity of the study in general as well as the complexity and variety of measures staff were required to administer. Staff demonstrated evidence of study and procedural knowledge on the study quiz as well as appropriate skill demonstration within the electronic data capture system and on six measures requiring certification. While the majority of staff had little to no experience in the six measures, all RAs demonstrated ability to correctly and reliably administer the measures throughout the trial.

There were several advantages of having delivered training in a staged manner using a variety of methods. For example, the extensive pre-training leading up to the training meeting led to trainees attending the training meeting with an equivalent knowledge base. Staff also asked advanced questions during the training meeting, many of which required further private discussion by the LT, ultimately improving study implementation. Trainings conducted via webinars with polling questions, as well as the personal response devices used during the training meeting, kept attendees engaged and provided immediate feedback regarding comprehension. Importantly, use of these interactive technologies alerted presenters to the need for additional training to ensure knowledge transfer and uptake. Finally, having centralized LT trainers responsible for certifying staff ensured standardized procedures across sites while being supported by local supervisors helping remediate weaknesses when necessary. Of note, difficulties RAs had in obtaining certification could generalize to other study related tasks, and staff members who had difficulty achieving certification often had difficulty in other areas, per discussions with site supervisors.

One major limitation of the current paper is that STRIDE was not designed to formally evaluate training procedures. As such, various training methods were not compared to other methods. Second, data was not systematically gathered to aid in calculating inter-rater reliability throughout the training, which could have supplemented the descriptions provided.

Several recommendations are made for trainers of multisite studies based on STRIDE experience. First, it is recommended that trainers not only assess scoring or rating abilities, but assess all skills required to accurately and reliably administer measures in trials, such as knowledge, interviewing skills, and documentation of data. These recommendations are similar to those outlined elsewhere (Del Boca, Babor, & McRee, 1994; Kobak et al., 2004) regarding essential skills and basic training recommendations yet it is not always possible to hire staff with these skills fully developed. As such, there is a need to assess and shape these behaviors, as was done in STRIDE. It is also recommended that protocol-specific training and evaluation be centralized in multi-site studies to ensure standardization across sites. For STRIDE, this was achieved by having two devoted LT trainers, with local supervisors providing additional inperson assistance when needed.

Second, it is recommended that trainers individualize the timeline for re-certifying staff. Studies often provide retraining on a generic timeline and/or re-evaluate research staff at 6 to 12 month intervals. While this approach may be suitable depending on the difficulty of the measures and the experience of the RAs, it is unlikely suitable for all staff. While the majority of STRIDE RAs demonstrated excellent maintenance of their skills at the 12 month recertification, a few staff members required retraining. The trainers observed those who had difficulty with re-certification also had difficulty with the initial certification. Therefore, it is recommended that trainers establish minimum re-certification timelines for all staff but individualize timelines based on each person’s initial performance. Thus, trainers may be required to more frequently work with staff to prevent drift and ensure skills are maintained, ultimately decreasing measurement error throughout the trial and ensuring outcomes are properly evaluated.

Third, it is recommended that trainers consistently provide feedback in a timely manner and be solely dedicated to the task of training during key study periods. At times during certification and re-certification phases in STRIDE, the trainers were unable to review all audio recordings received within a two week timeframe, a self-set maximum goal. To optimize trainees' skill development, constructive feedback needed to occur soon after the recording and before the trainee continued to practice incorrect behaviors. A system such as Rosen and colleagues'(2008) web-based interactive video system in which trainees individually administer measures to actors portraying various scenarios while centralized trainers evaluate the full complement of RA skills across sites can streamline the process of certification. Therefore, it is further recommended that technology be implemented to aid in the certification process when feasible.

Solicitation of past experience with study measures is recommended; however, it is not recommended that trainers strictly use information about staff experience as indicators of trainability or future performance. Sometimes STRIDE RAs who had more experience or education had greater difficulty meeting the training requirements. Similar to Targum's (2006) findings, staff competency was not achieved based on experience alone and staff improved their existing skills with training.

Overall, the staged training process with competency evaluations for STRIDE was comprehensive, iterative, and ensured all staff conducted study procedures and administered measures reliably. The techniques and recommendations described herein are easily replicable for others training staff for multisite studies.

Acknowledgments

We acknowledge and thank the numerous individuals who worked on STRIDE from the participating study sites (Gateway Community Services, Nexus Recovery Center, Gibson Recovery Center, Arapahoe House, Morris Village, Charleston VAMC, Memorial Hermann Prevention and Recovery Center, Penn Presbyterian Medical Center, and St. Luke's Roosevelt Hospital), sites' affiliated Regional Research and Training Centers (Florida Node Alliance, Texas Node, Ohio Valley Node, Southern Consortium Node, Delaware Valley Node, and the Greater New York Node), as well as the staff at the EMMES Corporation for serving as the Data and Statistics Center (DSC2) and the Clinical Coordinating Center (CCC).

FUNDING

Research reported in this publication was supported by the National Institute on Drug Abuse of the National Institutes of Health [Award Number U10DA020024 (PI: Trivedi)]. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

REFERENCES

1. Trivedi MH, Greer TL, Grannemann BD, Church TS, Somoza E, Blair SN, Nunes E. Stimulant reduction intervention using dosed exercise (STRIDE) - CTN 0037: Study protocol for a randomized controlled trial. Trials. 2011; 12:206. [PubMed: 21929768]
2. Baca-García E, Blanco C, Sáiz-Ruiz J, Rico F, Diaz-Sastre C, Cicchetti DV. Assessment of reliability in the clinical evaluation of depressive symptoms among multiple investigators in a multicenter clinical trial. Psychiatry Research. 2001; 102(2):163–173. [PubMed: 11408055]
3. Del Boca FK, Babor TF, McRee B. Reliability enhancement and estimation in multisite clinical trials. Journal of Studies on Alcohol, Suppl. 1994; 12:130–136.

4. Fussell H, Kunkel LE, McCarty D, Lewy CS. Standardized patient walkthroughs in the National Drug Abuse Treatment Clinical Trials Network: Common challenges to protocol implementation. American Journal of Drug & Alcohol Abuse. 2011; 37(5):434–439. [PubMed: 21854287]

5. Golden, CJ. Stroop Color and Word Test: A manual for clinical and experimental uses. Wood Dale: 1978.

6. Jeglic E, Kobak K, Engelhardt N, Williams JBW, Lipsitz JD, Salvucci D, Bellew K. A novel approach to rater training and certification in multinational trials. International Clinical Psychopharmacology. 2007; 22(4):187–191. [PubMed: 17519640]

7. Kobak KA. Inaccuracy in clinical trials: effects and methods to control inaccuracy. Current Alzheimer Research. 2010; 7(7):637–641. [PubMed: 20704557]

8. Kobak KA, Engelhardt N, Lipsitz JD. Enriched rater training using Internet based technologies: A comparison to traditional rater training in a multi-site 24 depression trial. Journal of Psychiatric Research. 2006; 40(3):192–199. [PubMed: 16197959]

9. Kobak KA, Engelhardt N, Williams JBW, Lipsitz JD. Rater Training in Multicenter Clinical Trials: Issues and Recommendations. Journal of Clinical Psychopharmacology. 2004; 24(2):113–117. [PubMed: 15206656]

10. Kobak KA, Lipsitz JD, Williams JBW, Engelhardt N, Bellew KM. A New Approach to Rater Training and Certification in a Multicenter Clinical Trial. Journal of Clinical Psychopharmacology. 2005; 25(5):407–412. [PubMed: 16160614]

11. Mulsant BH, Kastango KB, Rosen J, Stone RA, Mazumdar S, Pollock BG. Interrater Reliability in Clinical Trials of Depressive Disorders. American Journal of Psychiatry. 2002; 159(9):1598–1600. [PubMed: 12202285]

12. Rosen J, Mulsant BH, Marino P, Groening C, Young RC, Fox D. Web-based training and interrater reliability testing for scoring the Hamilton Depression Rating Scale. Psychiatry Res. 2008; 161(1): 126–130. [PubMed: 18760843]

13. Rush AJ, Trivedi MH, Ibrahim HM, Carmody TJ, Arnow B, Klein DN, Keller MB. The 16-Item Quick Inventory of Depressive Symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): A psychometric evaluation in patients with chronic major depression. Biological Psychiatry. 2003; 54:573–583. [PubMed: 12946886]

14. Sheehan DV, Lecrubier Y, Harnett Sheehan K, Janavs J, Weiller E, Keskiner A, Dunbar GC. The validity of the Mini International Neuropsychiatric Interview (MINI) according to the SCID-P and its reliability. European Psychiatry. 1997; 12(5):232–241.

15. Targum SD. Evaluating rater competency for CNS clinical trials. Journal of Clinical Psychopharmacology. 2006; 26(3):308–310. [PubMed: 16702896]

16. Wechsler, D. Wechsler test of adult reading. San Antonio: 2001.

17. WHO. Composite International Diagnostic Interview, version 2.1. Geneva: World Health Organization; 1997.

Table 1

Research measures administered by research assistants in STRIDE that required certification by centralized the Lead Investigator Team's trainers

Measure	Information Assessed	Skills Required
Quick Inventory of Depressive Symptomatology – clinician rated version (QIDS-C_{16})	Depression symptoms	– Ability to devise appropriate follow-up questions
Timeline Followback (TLFB)	Quantity and frequency of alcohol, nicotine, and drug use	– Ability to devise appropriate follow-up questions – Familiarity with drugs of abuse – General interviewing skills
Stroop Color and Word Test (Stroop)	Response inhibition	– Ability to follow fully standardized neurocognitive test administration requirements – Multitasking
Mini International Neuropsychiatric Interview (MINI)	DSM-IV Axis I disorders excluding substance abuse and dependence	– Semi-standardized question administration skills – Ability to devise appropriate follow-up questions – DSM-IV Axis I diagnostic familiarity
Wechsler Test of Adult Reading (WTAR)	Vocabulary (to estimate pre-morbid intelligence)	– Ability to follow fully standardized test administration requirements – Pronunciation familiarity
Stimulant Selective Severity Assessment (SSSA)	Stimulant withdrawal symptoms	– Ability to devise appropriate open-ended questions – Familiarity with symptoms and terminology (e.g., hypophagia)

Table 2

Summary of skills required by research assistants to achieve certification on measures,[a] problems identified during review of recordings submitted for certification review, and corrective actions taken by research assistants to achieve certification

Areas Assessed	Problems Identified	Feedback Provided and Corrective Actions Recommended to Achieve Certification
Scoring	– Inconsistent or inaccurate scoring – Incorrect use of Likert scale anchors when asking questions – Systematic under or over rating	– How to assess and calculate standard alcohol drinks, using specific example from recordings (TLFB) – Explicit explanation of how RA timed the task vs. how the task is to be timed (Stroop) – Examples of follow-up questions necessary for scoring (MINI) – Transcription of RA and participant statements with explicit description of errors and suggestions for improvement – Time stamp of problem area for RA to review on recording while reviewing written feedback – Supplemental information (WTAR: links to online dictionaries with audio pronunciations; MINI: diagnostic information)
Administration	– Poorly phrasing semi-structured or unstructured questions – Lack of thorough follow-up questions – Not reading fully standardized directions verbatim and following fully standardized instructions	– Transcription of poorly phrased questions, explanation/ rationale why phrasing was deemed poor, suggested alternative ways of asking questions in specific scenarios – Suggestions for appropriate follow-up questions and explanation of how gathering additional information impacts scoring – Transcription of verbatim directions/questions overlaid with deviations from the verbatim requirement and rationale why following standardization is important – Tips on how to implement multiple tasks (Stroop) – Request to review study manuals (page numbers provided) describing suggested question phrasing
Interviewing skills	– Question delivery rate too fast – Not allowing enough time for responses – Not using open ended questions – Few “customer service” and rapport building skills (e.g., encouraging, thanking participants for efforts)	– Explicit explanation of when speaking rate is too fast given the complexity of the directions/questions – Educating staff on cognitive effects of early abstinence from substance use and impact on ability to engage in interviews – Examples of open ended questions to use in specific scenarios – Suggestions for how to engage participants during standardized tasks and how to transition between measures – Request to review study manuals (page numbers provided) describing recommended ways to deal with difficult scenarios
Documentation	– Incorrect case report form completion – Incorrect conversion of visual analog scale marks to numerical equivalent – Error correction procedures not following Good Clinical Practices	– Clarification on expectations for case report form completion using specific examples of incorrect documentation and how it should be completed – Request to review study manuals (page numbers provided) describing how documentation is to be completed – Specific description of error correction problems and how it should be documented

[a] Timeline Followback (TLFB), Stimulant Selective Severity Assessment (SSSA), Quick Inventory of Depressive Symptomatology – clinician rated version (QIDS-C_{16}), Wechsler Test of Adult Reading (WTAR), and Stroop Color and Word Test (Stroop).