THE JOURNAL OF ALTERNATIVE AND COMPLEMENTARY MEDICINE
Volume 25, Number 11, 2019, pp. 1103–1108

DOI: 10.1089/acm.2019.0260

JACM

Can Reliability of the Chinese Medicine Diagnostic Process Be Improved? Results of a Prospective Randomized Controlled Trial

Rosa N. Schnyer, DAOM, LAc, IFMCP,[1] Patrick McKnight, PhD,[2] Lisa A. Conboy, MA, MS, ScD,[3,4] Eric Jacobson, PhD, MPH,[4,5] Anna T. Ledegza, ScD,[6] Mary T. Quilty, SM,[5] Roger B. Davis, ScD,[4] and Peter M. Wayne, PhD[5]

Abstract

Background: The diagnostic framework and clinical reasoning process of Chinese medicine are central to the practice of acupuncture and other related disciplines. There is growing interest in integrating it into clinical trials of acupuncture and Chinese herbal medicine to guide individualized treatment protocols and evaluate outcomes. Strategies that enhance diagnostic reliability may contribute to this integration.

Objectives: (1) To evaluate inter-rater reliability among practitioners of Traditional Chinese Medicine (TCM) when assessing women with dysmenorrhea using a structured assessment questionnaire (Traditional East Asian Medicine Structure Interview [TEAMSI]-TCM) compared to using a TCM questionnaire from routine clinical practice, not developed for research purposes (CONTROL); and (2) To evaluate the impact of training in the use of each approach on reliability.

Design: Thirty-eight acupuncturists were asked to complete assessments of 10 subjects based on the viewing of a videotape of the initial assessment interview, a picture of the tongue, and a description of the pulse. Acupuncturists were randomized into one of four groups comparing the use of two questionnaires, TEAMSI-TCM versus CONTROL, and comparing training in the use of each versus no training.

Analysis: The authors used Cohen's kappa to estimate agreement on TCM diagnostic categories relevant to dysmenorrhea between 2 practitioners with respect to questionnaires and training over all 10 patients and all 10 TCM diagnostic categories. For all analyses, the authors estimated kappa values for questionnaire, training, and experience level. Analysis of variance was used to test agreement among various groupings.

Results: Regardless of the questionnaire used or training, analysis of inter-rater reliability indicated overall agreement to be low among practitioners (median 0.26). Kappa varied slightly by questionnaire and training, among 38 practitioners, but the difference was not statistically significant ($p = 0.227$ and $p = 0.126$, respectively).

Conclusions: A structured assessment interview instrument designed for research purposes with or without training did not significantly improve reliability of TCM diagnosis of dysmenorrhea compared to a commonly used instrument. Challenges in assessing reliability in TCM remain.

Keywords: reliability, validity, traditional Chinese medicine, diagnostic methods

[1]Adult Health, University of Texas School of Nursing, Austin, TX.
[2]Department of Psychology, George Mason University, Fairfax, VA.
[3]New England School of Acupuncture, MCPH University, Worcester, MA.
[4]Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA.
[5]Osher Center for Integrative Medicine, Harvard Medical School and Brigham and Women's Hospital, Boston, MA.
[6]Alkermes, Inc., Waltham, MA.

Introduction

The personalized diagnostic framework and clinical reasoning process of Chinese medicine (CM)* are central to the practice of acupuncture and other related disciplines; it both guides and provides essential information about the therapeutic intervention. Integrating the Traditional Chinese Medicine† (TCM) diagnostic process in clinical studies of interventions based on TCM may be essential in increasing ecological validity and may help identify individual patient differences in response to treatment. Furthermore, delivering interventions that are individually tailored based on TCM pattern differentiation, in addition to biomedical diagnoses or symptoms (e.g., back pain and hypertension), could potentially enhance precision in delivering personalized treatment, resulting in improved clinical outcomes. The inclusion of a TCM diagnosis as a feature in clinical trials targeting biomedical conditions, however, will require that reliability and validity of the clinical reasoning process can be established.

Reliability provides an estimation of the precision with which a diagnosis can be consistently and repeatedly obtained by different practitioners (inter-rater) under the same conditions or by the same practitioner under different conditions (intra-rater). Validity refers to the general accuracy of the diagnostic conclusions and represents the best available approximation to the right diagnosis.[1] Increasingly, rigorous attempts have been made to evaluate the reliability of the CM diagnostic framework in the context of clinical research[2,3]; methodological issues have limited interpretation of many of these studies, and other than a few exceptions, overall agreement has been modest.[2–4] Various factors have been cited to account for the poor reliability found in many of these studies. Their overarching goal was to address three key challenges often cited in assessing reliability in TCM: (1) divergence in the diagnostic reasoning process and recoding of findings; (2) variations in the level of training and experience of practitioners; and (3) discrepancy in the information given by patients to different providers at different points in time. The authors designed a prospective randomized controlled trial aimed at assessing whether the use of a structured instrument developed for research purposes (to address challenge #1), in combination with training (to address challenge #2), could increase inter-rater reliability of the TCM diagnostic process based on recorded interviews (to address challenge #3). Their aim was *not* to assess reliability in a naturalistic clinical setting, but rather to explore innovative methods of increasing reliability in clinical trials of TCM with the purpose of making better sense of diagnostic data collected in these studies. This report presents the results of a prospective randomized controlled trial study designed specifically to improve reliability for research purposes.

Developing and validating assessment instruments

In a previously published article, the authors described the process of developing a structured assessment interview for TCM[5] TEAMSI-TCM,‡ specifically developed with data collection in mind. The purpose was to produce a validated structured interview questionnaire that reflects the clinical reasoning process of TCM, captures diagnostic data consistently and reliably, and improves inter-rater reliability of TCM diagnosis. TEAMSI-TCM is a *prescriptive* instrument meant to guide clinicians to use the proper indicators, combine them in a systematic manner, and generate conclusions; it was designed for use in combination with training. The authors developed this version of TEAMSI-TCM to be used in dysmenorrhea and related women's health conditions and validated and tested it by comparing it to a descriptive questionnaire commonly used in clinical practice, the NESA (New England School of Acupuncture at MCPH University) clinic TCM questionnaire (herein referred to as CONTROL), which is similar to many questionnaires used by TCM-trained practitioners. In addition, the authors evaluated the impact of training on improving reliability and validity.

The Institutional Review Boards of the New England School of Acupuncture and Harvard Medical School approved the study. Patients and practitioners signed informed consent. Clinicians were compensated for their time; patients received a gift certificate for their participation.

Methods

Licensed clinicians ($n=38$) who had practiced TCM style acupuncture for at least 3 years were recruited from the greater Boston area to complete a TCM assessment of 10 patients with dysmenorrhea by watching videotaped clinical interviews. Their 2×2 design randomized clinicians to one of four groups to use one of two assessment questionnaires and to receiving or not receiving training in the use of the respective questionnaire (Fig. 1). The number of practitioners (38) provided roughly equal number of assessment ratings per group; the number of patient assessments (10), aimed at providing variance of severity and clinical presentations; numbers were not based on power calculation.

Clinical interviews

Videotaped clinical interviews of 10 patients with self-reported dysmenorrhea (mean age: 29; mean pain severity on a numerical rating scale 1–10: 8; $n=7$ moderate; $n=3$ severe) were performed by an outside senior practitioner (author R.N.S.) without the use of either questionnaire (TEAMSI-TCM vs. CONTROL) and were augmented by a professional picture of the tongue and a record of the pulse taken at the time of the original assessment. Patients were recruited from the Boston area through web-based ads using an online bulletin board (Craigslist) and were first asked to complete the patient component of the two questionnaires; first the CONTROL followed by TEAMSI-TCM.

Training

Practitioners were provided an overview of the project and were trained in the use of their assigned questionnaire,

*The term CM is used here to include any discipline derived from the basic body of knowledge of CM.

†TCM refers specifically to the style of diagnosis and treatment based on Eight Principle Pattern Differentiation model.

‡The term Traditional East Asian Medicine (TEAM) replaces the term Oriental medicine. TEAMSI-TCM stands for Traditional East Asian Medicine Structure Interview, based on TCM.

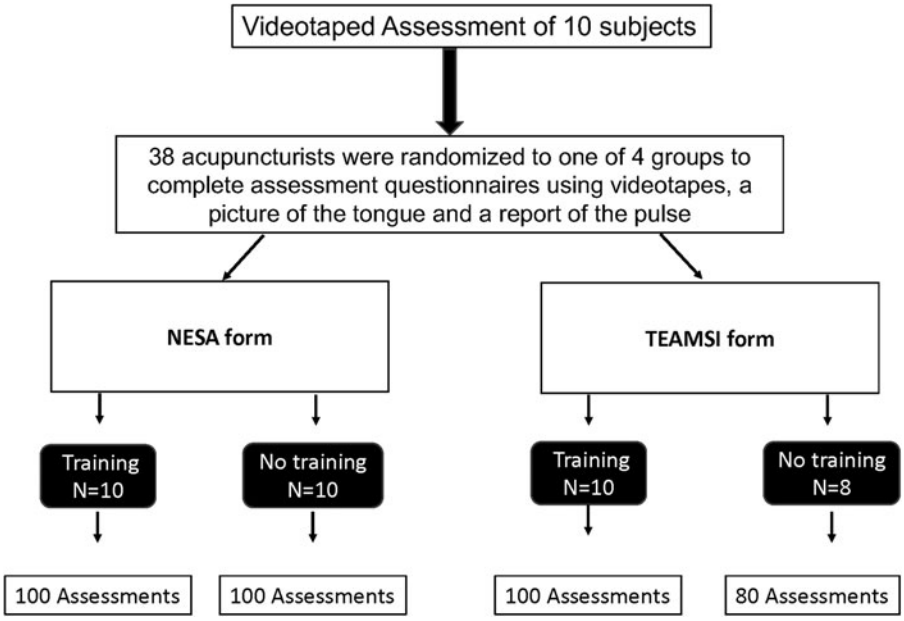

FIG. 1. Study design.

TEAMSI-TCM or CONTROL. Training in the TEAMSI-TCM arm lasted 8 h to familiarize practitioners with the use of this new questionnaire and to practice a structured interview process guided by the TEAMSI-TCM user's manual. The training session in the use of the CONTROL lasted 2 h. Both groups received training on how to use previously collected data (videotaped assessments, tongue and pulse), to consistently complete their assigned questionnaire. Practitioners were instructed to refrain from discussing the assessments with other participating practitioners to minimize discussion led agreement.

Testing

The clinical testing took place over two separate days and required ~14 h. Groups of 3–10 practitioners ranging in clinical experience (average years of experience given in parentheses) watched 10 individual patient videos (45–60 min each) and were given time to complete a clinical evaluation of each patient (30–60 min) using the questionnaire assigned to them: (a) TEAMSI with training (14.6 years); (b) TEAMSI, no training (7.1 years); (c) NESA with training (6.5 years); and (d) NESA, no training (12.3 years), Figure 1. To guard against fatigue, practitioner participants took regular breaks, and refreshments were provided. Qualitative interviews and surveys were conducted with the practitioners who participated in the study to assess face and content validity of the questionnaires and ecological validity of the testing processes (Conboy L, Schnyer R, Shaw J, McCallister A. Qualitative assessment of a reliability and validity testing in Traditional Chinese Medicine. Manuscript in preparation.). Two separate sets of packages were prepared for each individual patient with the corresponding TEAMSI or CONTROL questionnaires, including tongue and pulse information.

Testing day 1. After viewing all videos clinicians were asked to (1) provide in order of priority their own set of TCM diagnostic categories or patterns for each patient (without being prompted to select from a list); (2) state their own final TCM assessment (combination of patterns); and (3) present the treatment principles indicated to address each pattern. The authors first collected a practitioner generated set of TCM patterns because the authors wanted to capture naturalistically the clinical findings of each practitioner, independent of exposure to structured pattern differentiation in the data collection forms (Supplementary Appendix SA1).

Testing day 2. After viewing all videos and completing their own selection of TCM patterns for each patient, practitioners were asked to: (1) go back and review the materials for each patient (not including their own TCM assessment); and (2) using a *structured* ''pattern differentiation form'' (Supplementary Appendix SA1), select from a list of 10 basic patterns those presented by each patient. Their team selected these patterns from three CM textbooks[6–8] and identified them as being consistently and commonly encountered in dysmenorrhea (Table 1). This ''Pattern Differentiation Form'' (Supplementary Appendix SA1) provided a common framework on which to assess reliability and validity of TCM ''diagnosis'' across all groups independent of questionnaire (TEAMSI-TCM questionnaire or CONTROL questionnaire) or training. This structured data collection form allowed us to standardize the names of the patterns to address divergence in the recoding of findings (challenge #1). Given the multiple translations of CM terms (e.g., vacuity vs. deficiency), the authors wanted to standardize nomenclature to avoid confusion. Furthermore, the authors wanted to avoid assuming absence of

Table 1. Dysmenorrhea (Simple) Patterns

Repletion	*Vacuity*
1. Qi stagnation	7. Qi vacuity
2. Blood stasis	8. Blood vacuity
3. Stagnation transforming into fire	9. Yang vacuity
4. Accumulation of cold	10. Yin vacuity
5. Accumulation of cold dampness	
6. Damp heat	

patterns presence if they were not endorsed and capture the rationale used by practitioners to endorse the presence of a specific pattern.

Assessing reliability and validity of TCM "diagnoses"

To arrive at their final selection in the ''pattern differentiation'' form, clinicians were asked to complete a TCM assessment for each patient by: (1) indicating the *pattern's presence* (1 = present, 0 = absent) and to state the rationale for this selection (i.e., etiology, tongue, pulse, and so on); (2) indicating if each specific pattern was the *''primary'' pattern*; and (3) rating on a 1–10 numerical analog scale (1 = minimally, 10 = absolutely) how much each pattern characterized this particular patient's clinical presentation (*clinical relevance*).

Reliability and validity (Conboy L, Schnyer R, Shaw J, McCallister A. Manuscript in preparation.) were assessed on the final TCM ''diagnosis'' (i.e., the *pattern's presence*) *only*. Their first aim was to assess inter-rater reliability, the degree of agreement between practitioners on the TCM pattern differentiation when assessing the same patient. In this particular study the authors wanted to assess inter-rater reliability among a broad spectrum of TCM trained practitioners in a limited number of patients (i.e., women with dysmenorrhea). Their secondary aims were to evaluate inter-rater reliability of TCM patterns in dysmenorrhea when using a systematically developed assessment instrument (TEAMSI-TCM) and to evaluate if training—on how to use this instrument—affected reliability.

Analysis

Their main objective was to assess inter-rater reliability by form (TEAMSI-TCM vs. control) and Training (in structured interview using each form) versus no training. In addition, the authors conducted other exploratory analyses that were not likely to be statistically powered to show differences (Table 2).

Inter-rater reliability. Cohen's kappa[9] was used to estimate agreement on TCM diagnostic categories (Table 1) relevant to dysmenorrhea (''pattern presence'') between two clinicians (*inter-rater reliability*) with respect to questionnaires and training across all 10 patients and all 10 TCM diagnostic categories (day 2). For all analyses, kappa values were estimated for questionnaire, training, and experience level across all patterns. The authors used the conventional method for calculating kappa without weighting. Kappa values were averaged. To test whether group differences (e.g., questionnaire, training, and experience level) affected agreement, the authors used a one-way analysis of variance (ANOVA) *F*-test.

Although Cohen's kappa test was not designed for multiple binary variables analyzed across multiple raters, at the time of this study, it remained the standard and simplest statistic reported in the medical diagnostic literature. The authors chose Cohen's Kappa (rather than Fleiss Kappa) because their intent was to do all possible comparisons and identify if any two raters were substantially different. Similarly, although ANOVA is rarely indicated for models where data are not independently and identically distributed, the authors chose to report only the simplest statistics. The authors analyzed the data with better-suited analyses when indicated and found that the results of more complex statistical models concurred with this simpler one (data available upon request).

Table 2. Types of Analysis Used and Purpose

Metric	*Purpose*
Kappa	To estimate agreement on TCM diagnostic categories relevant to dysmenorrhea (''pattern presence'')
Signal detection parameters (sensitivity, specificity, efficiency)	To determine how well the combination of either form or training allows TCM practitioners discern the correct patterns
ANOVA	To test whether group differences (e.g., form, training, experience level) affected agreement
Endorsement analysis	To better inform the interpretation of the agreement analysis, the authors first observed the frequency with which the experts, relative to experience, form, and training, endorsed each of the 10 patterns.

ANOVA, analysis of variance; TCM, Traditional Chinese Medicine.

Face, content, and ecological validity. To assess whether TEAMSI-TCM was consistent with the established Pattern Differentiation methodology characteristic of TCM, that it covered a full range of TCM assessment methods, and that it was in fact usable by acupuncturists, the authors performed a qualitative analysis of semistructured interviews conducted with 27 of the practitioner participants.[4] Practitioners were asked to comment on the questionnaires that they were assigned to use (TEAMSI-TCM vs. NESA) and were specifically asked about the questionnaires in terms of face, content, and ecological validity. A complete description of the face, content, and ecological validity testing is not included in this report (published elsewhere [Conboy L, Schnyer R, Shaw J, McCallister A. Manuscript in preparation.]).

Naturalistic versus prompted clinical findings. The results reported below are based on analysis of agreement of naturalistic clinical findings obtained during day 1 of testing, which were recorded in the practitioners' handwriting. An independent acupuncturist, who did not participate in the study and who was blind to assessor, patient, and questionnaire, transferred these findings onto the same structured Pattern Differentiation Form used to prompt practitioner responses during day 2, to compare them. All the same analyses of agreement were then conducted in the Pattern Differentiation Form. The authors found no significant differences in the results.

Results

Inter-rater reliability (agreement among 38 experts)

The overall within-group kappa estimated among the 38 clinicians (i.e., same questionnaire and training group) was 0.26 (mean). Concordant kappa values varied by questionnaire and training (see shaded diagonal in Table 3) with the CONTROL questionnaire group with training showing the highest relative agreement among the clinicians. The off

Table 3. Inter-Rater Reliability: Agreement Among 38 Clinicians

		TEAMSI		*NESA*	
		Training	*No training*	*Training*	*No training*
TEAMSI	Training	0.21			
	No Training	0.26	0.26		
NESA	Training	0.28	0.29	0.31	
	No Training	0.23	0.27	0.29	0.26

Mean kappa values across all clinicians within each group. SE for all four groups were equivalent (after rounding up) at 0.03.

Shaded values indicate if our inter-rater agreement was better, the same, or worse than chance. Kappa is a measure of this difference, standardized to lie on a −1 to 1 scale. One would indicate perfect agreement, zero is exactly what would be expected by chance, and negative values indicate agreement less than chance. Our found values or 0.21 to 0.31 indicate that our agreement was better than chance, but far from strong agreement.

NESA, New England School of Acupuncture; SE, standard error; TEAMSI, Traditional East Asian Medicine Structure Interview.

diagonal kappa values (not shaded) in Table 3 indicate the between-group agreement; the between-group kappa values were approximately the same as the within-group kappa values (mean 0.26 for concordant groups and 0.27 for nonconcordant groups). These results indicate poor overall agreement regardless of the questionnaire or training.

Between-group reliability estimates

Testing of kappa values by group indicated no significant differences by experience levels (<10 and >10 years); the median kappa for both groups was equal to the median overall kappa at 0.29. However, experience did appear to modify the kappa observed across questionnaire and training; the highest median kappa was observed in the more experienced CONTROL/training group (0.41). The lowest median kappa was observed in the more-experienced/TEAMSI/training group (0.24). The groups with less experienced raters, who received no training, performed similarly, regardless of whether they used the CONTROL questionnaire (0.33) or the TEAMSI questionnaire (0.32).

Frequency of pattern endorsement

To better inform the interpretation of the agreement analysis, the authors first observed the frequency with which the experts, relative to experience, questionnaire, and training, endorsed each of the 10 patterns. The endorsement analysis revealed that everyone endorsed patterns in similar relative frequencies (Table 4). Kappas were symmetric in each group (means and medians coincided) and the variability remained relatively consistent by group; therefore, testing of agreement among various groupings was done using ANOVA.

The one-way ANOVA indicated that the groupings (questionnaire, training, experience) did not significantly explain the variation in kappa: by questionnaire and training (among 38 raters $p=0.13$, compared to expert $p=0.23$); by level of experience (among 38 raters $p=0.83$, compared to expert $p=0.88$); or by experience, questionnaire, and training (among 38 raters $p=0.14$, compared to expert $p=0.13$). Looking at individual predictors of each model (using a Likelihood Ratio Test of nested models) the authors found some evidence that the interaction of questionnaire and training ($p=0.05$) may have driven the ascent/descent of the kappa values and that experience level may have modified questionnaire/training effects ($p=0.02$). None of the results differs when analyses were performed on the structured ''Pattern Differentiation Form'' (day 2). The most frequently endorsed patterns among all 38 raters were: #1 Qi stagnation, #2 Blood Stasis, and #7 Qi Vacuity.

Face, content, and ecological validity

TEAMSI scored higher on face and content validity than CONTROL, but lower on usability (results reported elsewhere [Conboy L, Schnyer R, Shaw J, McCallister A. Manuscript in preparation.]). Most of the practitioners reported that a few aspects of the testing procedure inhibited their ability to perform their diagnosis well and that it impaired the establishment of therapeutic rapport. The primary hindrance was a lack of familiarity with the TEAMSI questionnaire.

In addition, the authors estimated signal detection parameters, including sensitivity, specificity, and efficiency,[10] as well as face validity, data not included in this report.

Table 4. Frequency of Pattern Endorsement by Groups: Experience, Form, and Training

Pattern	*Gold standard*	*Overall*	*TEAMSI*	*NESA*	*Training*	*No training*	*More experienced >10 years*	*Less experienced <10 years*
1	1.00	0.87	0.96	0.80	0.91	0.84	0.82	0.91
2	1.00	0.73	0.76	0.74	0.73	0.74	0.83	0.73
3	0.70	0.30	0.28	0.32	0.33	0.26	0.32	0.33
4	0.80	0.24	0.34	0.16	0.25	0.24	0.27	0.25
5	0.60	0.19	0.26	0.12	0.19	0.18	0.18	0.19
6	0.30	0.27	0.33	0.22	0.22	0.32	0.24	0.22
7	1.00	0.71	0.74	0.68	0.70	0.72	0.73	0.70
8	0.80	0.40	0.48	0.34	0.47	0.34	0.41	0.47
9	0.10	0.21	0.29	0.14	0.23	0.18	0.26	0.23
10	0.50	0.36	0.46	0.26	0.34	0.36	0.34	0.34

Discussion

Overall inter-rater reliability was poor as indicated by the low kappa values. Differences by questionnaire and training were not significant. Several potential limitations of their study might adversely affect their estimates of agreement. The CM diagnostic assessment process involves an interactive, comprehensive history taking; each sign and symptom has a contextual rather than a categorical clinical application; the significance and meaning of a particular symptom, for example, dysmenorrhea, change depending on the context. Due to the constraints of this experimental design, this context-dependent relationship was altered. Clinicians did not get to directly interact with patients, but rather used the instruments to rate patients by observing a videotaped interview, looking at a picture of the tongue, and relying on a report of the pulse findings. Why did the authors opt for this design? The authors aimed to decrease discrepancy in the information given by patients to different providers at different points in time and to reduce the artifact created by order in which raters assessed participants (challenge #3).

In the qualitative interviews, clinicians reported that TEAMSI was thorough and comprehensive (good face and content validity), but was difficult to use (resulting in poor ecological validity). Trained clinicians expressed a need for more training; those using TEAMSI who were not trained reported that the questionnaire was too foreign to be comfortably used diagnostically (Conboy L, Schnyer R, Shaw J, McCallister A. Manuscript in preparation.). Consequently, trained clinicians may have relied on familiar behaviors—skipping through the structured process and completely bypassing the complexity of the questionnaires—and in this way more accurately identifying the patterns and therefore increasing reliability. It is possible that training was inadequate, thus creating a deficit by instilling doubt in the clinicians about their diagnostic abilities and interfering with the process of accurately assessing patients. It is also important to note that practitioners were not randomized into each group equally by average years of experience, which may have influenced the results.

All diagnostic patterns were treated equally; some patterns, however, may be more relevant to dysmenorrhea. A subsequent analysis weighting pattern selected *a priori* in terms of diagnostic importance might provide a more useful metric of agreement. One major limitation with kappa is the sensitivity to extreme base rates, and their results reflect several extreme base rates for both high and low values across the 10 patterns. While these extreme cases might usually affect the overall kappa, they did not in this case. Although the authors selected a heterogenous patient sample in terms of severity of symptoms, the patients may have been too homogenous. Their findings are limited to this condition and patient population and might not be generalizable to other conditions.

Their study is innovative by comparing an instrument commonly used in clinical practice with an instrument developed for research purposes, by assessing the potential impact of training on reliability and using various quantitative and qualitative measures. Yet, it highlights the challenge of designing a study of reliability in TCM diagnosis while addressing all potential limitations. Future work will require assessing reliability in an interactive setting, simplifying the structured interview while maintaining face and content validity, and improving training for practitioners to gain sufficient familiarity with the instruments and the experimental process.

Acknowledgments

The authors thank all the participating acupuncturists. The authors also thank Monica Shields, Ellen Connors, and Andrea Hrbek, for their assistance with this project.

Disclaimer

Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NCCAM or the National Institutes of Health.

Author Disclosure Statement

No competing financial interests exist.

Funding Information

This work was supported by National Center for Complementary and Integrative Health (NCCIH) Grant Numbers 5 U19 AT002022, K24 AT009282.

Supplementary Material

Supplementary Appendix SA1

References

1. Cook T, Campbell D. Quasi-Experimentation: Design and Analysis Issues for Field Settings. Chicago: Rand Mc Nally, 1979.
2. Jacobson E, Conboy L, Tsering D, et al. Experimental studies of inter-rater agreement in Traditional Chinese Medicine: A systematic review. J Alternat Complement Med 2019;25:1085–1096.
3. O'Brien K, Birch S. A review of the reliability of traditional East Asian medicine diagnoses. J Altern Complement Med 2009;15:353–366.
4. Mist S, Ritenbaugh C, Aickin M. Effects of questionnaire-based diagnosis and training on inter-rater reliability among practitioners of Chinese Medicine. J Altern Complement Med 2009;15:703–709.
5. Schnyer R, et al. Development of a Chinese medicine assessment measure: An interdisciplinary approach using the Delphi Method. J Complement Altern Med 2005;11:1005–1013.
6. Flaws B. My Sister The Moon. Boulder: Blue poppy Press, 1992.
7. Maciocia G. Obstetrics and Gynecology in Chinese Medicine. London: Churchill Livingstone, 1998.
8. Wiseman N, Feng Y. A Practical Dictionary of Chinese Medicine. Brookline: Paradigm Publications, 1998.
9. Fleiss J. Measuring nominal scale agreement among many raters. Psychol Bull 1971;76:378–382.
10. Swets J. Signal Detection Theory and ROC Analysis in Psychology and Diagnostics. London: Psychology Press, 1996.

Address correspondence to:
Rosa N. Schnyer, DAOM, LAc, IFMCP
University of Texas at Austin, School of Nursing
1710 Red River St
Austin, TX 78712

E-mail: rschnyer@mail.nur.utexas.edu