THE JOURNAL OF ALTERNATIVE AND COMPLEMENTARY MEDICINE
Volume 15, Number 7, 2009, pp. 703–709

DOI: 10.1089/acm.2008.0488

Original Articles

Effects of Questionnaire-Based Diagnosis and Training on Inter-Rater Reliability Among Practitioners of Traditional Chinese Medicine

Scott Mist, Ph.D., L.Ac.,[1] Cheryl Ritenbaugh, Ph.D., M.P.H.,[2] and Mikel Aickin, Ph.D.[2]

Abstract

Objective: To investigate whether a training process that focused on a questionnaire-based diagnosis in Traditional Chinese Medicine (TCM), and developing diagnostic consensus, would improve the agreement of TCM diagnoses among 10 TCM practitioners evaluating patients with temporomandibular joint disorder (TMJD).
Design and setting: Evaluation of a diagnostic training program at the Department of Family and Community Medicine, University of Arizona, Tucson Arizona, and the Oregon College of Oriental Medicine, Portland, Oregon.
Subjects: Screened participants for a study of TCM for TMJD.
Practitioners: Ten (10) licensed acupuncturists with a minimum of 5 years licensure and education in Chinese herbs.
Methods: A training session using a questionnaire-based diagnostic form was conducted, followed by waves of diagnostic sessions. Between sessions, practitioners discussed the results of the previous round of participants with a focus on reducing variability in primary diagnosis and severity rating of each diagnosis: 3 waves of 5 patients were assessed by 4 practitioner pairs for a total of 120 diagnoses. At 18 months, practitioners completed a recalibration exercise with a similar format with a total of 32 diagnoses. These diagnoses were then examined with respect to the rate of agreement among the 10 practitioners using inter-rater correlations and kappas.
Results: The inter-rater correlation with respect to the TCM diagnoses among the 10 practitioners increased from 0.112 to 0.618 with training. Statistically significant improvements were found between the baseline and 18 month exercises ($p < 0.01$).
Conclusions: Inter-rater reliability of TCM diagnosis may be improved through a training process and a questionnaire-based diagnosis process. The improvements varied by diagnosis, with the greatest congruence among primary and more severe diagnoses. Future TCM studies should consider including calibration training to improve the validity of results.

Introduction

Whole systems research is theoretically congruent with the medicine being investigated.[1–8] It can examine the smallest portion of a medicine—does a treatment work for a specific symptom in a population—or the broadest research questions, such as examining the role of Traditional Chinese Medicine (TCM) within the biomedical system in the United States. However, each of these questions must take into account the theoretical underpinnings of the modality under investigation to be considered whole systems.

Too often this is where complementary and alternative medicine (CAM) research encounters problems. For example, there have been several recent studies of the effectiveness of a single acupuncture point for the treatment of complex biomedical conditions.[9–12] It is never the case that a trained TCM practitioner would use the same treatment for all patients with these conditions, nor would they use a single point. From the perspective of the TCM practitioner, this does not make theoretical sense.[13,14] An example from biomedicine illustrates the problem. In many Chinese hospitals, patients ask antibiotics for all sorts of conditions. Researchers

[1]Oregon College of Oriental Medicine, Portland, OR.
[2]Department of Family and Community Medicine, The University of Arizona, Tucson, AZ.

would conclude that antibiotics do not work if they studied them for flu or earaches without understanding that these conditions can have multiple causes. Whole systems research attempts to address this theoretical mistake, by remaining attuned to the theoretical basis of each system.

As CAM studies move towards a whole systems approach, the role of diagnosis within the system being investigated becomes increasingly important. The difficulty within TCM is that the existing studies have shown poor reproducibility between practitioners.[15–17] This should not be used to cast doubt about the validity and/or role of diagnosis within TCM. Biomedical diagnoses suffer from similar difficulties,[18–20] as do psychological diagnoses.[21,22]

The current study investigated whether practitioners could be trained to diagnose with greater inter-rater reliability. This is an important methodological step toward investigating whether TCM diagnosis is an important part of the success or failure of the TCM treatment patients receive.

Ten (10) TCM practitioners were participating in a larger whole systems study of TCM treatments for temporomandibular joint disorder (TMJD) in Tucson, Arizona, and Portland, Oregon. At each site there were 4 treating TCM practitioners and 1 expert diagnosing TCM practitioner. The study protocol called for a diagnosis at initial recruitment and one year later by a diagnosing TCM practitioner not involved in treatment, and at every TCM treatment by the treating practitioner. In order to prepare for the study, all TCM practitioners from Tucson and Portland participated in a joint calibration session in Tucson prior to study start, and then participated in local recalibration exercises 18 months later.

Materials and Methods

For the purposes of this study, calibration refers to the process of moving the practitioners toward the same TCM diagnosis when interviewing the same patient. Recalibration was a follow-up exercise in which we completed a second round of patient interviews to estimate reliability among practitioners. If any drift in diagnostic styles occurred, a similar training cycle would be implemented. Reliability, however, is a separate concept and refers to the consistency of two or more practitioners producing the same diagnosis for the same patient.

In a previous study, 10 TCM practitioners diagnosed participants using an open-ended questionnaire designed to cover the major distinguishing factors of diagnoses found in TMJD patients.[23] The questionnaire allows for multiple diagnostic outcomes, reflecting common practice among TCM practitioners. The form was designed to follow the usual TCM diagnostic interview process, beginning with the chief complaint and then identifying the key components of the TCM 10 Questions. The tongue, pulse, and observations are recorded, followed by the organs most affected. At the end of the form, 19 diagnoses are each given a score of 0 to 10 to indicate their severity or clinical relevance; space is provided to write in additional diagnoses. In this way the practitioners are guided from a broad perspective towards a diagnosis at the end of the form. Our form was designed to organize the vast amount of health history and current complaints into a format that would assist the practitioner to assess and rank the diagnoses. Figure 1 shows the final page of the diagnostic questionnaire.

Two-and-a-half days were set aside for the initial practitioner calibration. The first half-day focused on familiarization with the questionnaire-based diagnostic form, the study population (noting common diagnoses and presentations from the previous study[23]), and the treatment protocol. There were also discussions about some of the key characteristics that practitioners use to distinguish among diagnoses, including the overall importance of tongue and pulse observations.

On the second day, practitioners were paired for the diagnostic sessions. Practitioners were paired with each other in a round robin manner, rotating with each new participant. A total of 15 participants were interviewed by four pairs of practitioners for a total of 120 diagnoses. Women and men were recruited in Tucson by newspaper advertisements and flyers at local medical offices. The participants' complaints ranged from healthy to complex cases with multiple chronic diseases.

For each session, one practitioner would lead the interview while the other would simply take notes on the inquiry. After the lead interviewer exhausted the inquiry portion of the diagnosis, the second practitioner would ask any additional questions which were felt to be necessary to clarify the diagnosis. Then both practitioners would assess pulse and tongue individually. Finally, they would review their own notes and score the diagnoses. The practitioners were not allowed to discuss the case until the process was finished. After each patient was interviewed, questionnaires with diagnosis scores were returned to one of the authors, who data-entered them immediately for subsequent real-time review.

Between waves of participants, the entered scores were shown to the whole group of practitioners, with practitioners able to see their own scoring in relation to all the others. Practitioners, led by the first author, discussed any outlying diagnoses or severity scores. These discussions were intended to help the practitioners develop convergence on future diagnoses as well as on the meaning of the severity scores.

A similar process was completed 18 months later at both sites to evaluate how well the training was maintained. The recalibration took place in three sessions: with practitioners in Tucson, practitioners in Portland, and 2 diagnosticians. All interviews happened in pairs as in the calibration exercise, with partners rotating among each other, and within pairs switching the lead interviewer. After two diagnostic sessions, practitioners reviewed the outcomes. Participants who were being diagnosed had TMJD signs or symptoms but had not yet received a diagnosis of TMJD.

The diagnosticians also had an additional exercise in which they each made diagnostic scores based only on diagnostic questionnaires already completed by practitioners on study participants. This was done to begin to understand whether the information that is captured by the form is adequate for diagnosis and, if not, what additional information is needed. They reviewed the data collected from 11 questionnaires without diagnosis to determine if the form contained enough information from which to make a diagnosis.

Statistical analysis

The main analysis uses the Fleiss' kappa statistic,[24] a measure for assessing the chance-corrected agreement between a set number of raters when assigning categorical ratings to a

Diagnosis:
1. Liver *Qi* Constraint
0 1 2 3 4 5 6 7 8 9 10
2. Liver Blood *Xu*
0 1 2 3 4 5 6 7 8 9 10
3. Liver *Yin Xu*
0 1 2 3 4 5 6 7 8 9 10
4. Liver Wind
0 1 2 3 4 5 6 7 8 9 10
5. Liver *Yang* Rising
0 1 2 3 4 5 6 7 8 9 10
6. Liver Heat **Liver Fire Blazing**
0 1 2 3 4 5 ☐ 6 7 8 9 10
7. *Qi* and Blood Stagnation
0 1 2 3 4 5 6 7 8 9 10
8. Heart *Xu*: *Yin* / Blood / *Qi* / *Yang* **(Circle one. If more than one enter in Other)**
0 1 2 3 4 5 6 7 8 9 10
9. Spleen *Qi Xu*
0 1 2 3 4 5 6 7 8 9 10
10. Spleen Damp
0 1 2 3 4 5 6 7 8 9 10
11. Kidney *Yang Xu*
0 1 2 3 4 5 6 7 8 9 10
12. Kidney *Qi Xu*
0 1 2 3 4 5 6 7 8 9 10
13. Kidney *Yin Xu*
0 1 2 3 4 5 6 7 8 9 10
14. Kidney *Jing Xu*
0 1 2 3 4 5 6 7 8 9 10
15. Damp Heat Retention
0 1 2 3 4 5 6 7 8 9 10
16. Wind-cold invasion
0 1 2 3 4 5 6 7 8 9 10
17. Other: _____
0 1 2 3 4 5 6 7 8 9 10
18. Other: _____
0 1 2 3 4 5 6 7 8 9 10

FIG. 1. Scaled diagnoses in diagnostic questionnaire.

number of items,[24,25] providing overall agreement by wave of participants. Fleiss' kappa, unlike the more familiar Cohen kappa, is appropriate when there are multiple raters. This is the most conservative approach; however, this methodology does not reflect the realities of the TCM diagnostic system.

Most participants had multiple diagnoses; only one practitioner diagnosed a participant as having a single diagnosis in the calibration exercise. Further, within the field of TCM diagnostic classifications, a natural progression of disease is observed. For example, one practitioner might feel that a participant has severe Liver Blood *Xu* while the next might diagnose the participant as having Liver *Yin Xu*. In TCM, Blood *Xu* leads to *Yin Xu* and *Qi Xu* leads to *Yang Xu*. Therefore, a second kappa, called whole systems kappa for clarity, was calculated; this counted agreement as including both those patients whose top two diagnoses were the same, and those whose diagnoses differed only based on disease progression, as described above. This analysis counted those cases in which the practitioners agreed on the top two diagnoses but disagreed on which was the most severe as agreement. Cases where the practitioners identified the same organ but differed on the substance (e.g., Liver Blood *Xu* and Liver *Yin Xu*) were considered in agreement. Likewise, Heart Blood *Xu* and Heart *Yin Xu*; Heart *Qi Xu* and Heart *Yang Xu*; and Kidney *Qi Xu* and Kidney *Yang Xu* were considered to be in agreement. This effectively reduced the number of diagnoses present by four.

Finally, intraclass correlation, using a two-way mixed effects model in which practitioner effects are random and diagnosis effects are fixed, was used to compare agreement on the most common diagnoses. Only those cases where practitioners diagnosed the participants with the same primary diagnosis (the diagnosis with the highest severity rating) were considered as matching.

Table 1. TCM Diagnosis in Calibration Study and Current TCM for TMD Study Population

	Calibration population (N = *82*)			*TCM for TMD Study population* (n = *168*)		
Diagnoses	N *Percent*	*Average severity* *StDev*	*Range*	N *Percent*	*Average severity* *StDev*	*Range*
Liver *Qi* Constraint	72 87.8%	4.6 1.9	1–10	146 86.4%	5.0 2.0	1–9
Qi & Blood Stagnation	58 70.7%	4.7 1.7	1–9	138 81.7%	5.6 1.6	2–9
Spleen *Qi Xu*	55 67.1%	3.6 1.6	1–8	107 63.3%	3.6 1.5	1–7
Liver Blood *Xu*	20 24.4%	2.8 1.3	1–6	67 39.6%	3.4 1.5	1–7
Kidney *Yin Xu*	33 40.2%	3.2 1.8	1–10	66 39.1%	2.9 1.5	1–8
Kidney *Qi Xu*	23 28.0%	2.7 1.7	1–8	62 36.7%	2.7 1.3	1–8
Liver *Yin Xu*	24 29.3%	3.0 1.4	1–6	44 26.0%	3.0 1.4	1–6
Liver Heat	24 29.3%	3.2 1.8	1–9	38 22.5%	3.0 1.4	1–6
Heart *Yin Xu*	15 18.3%	3.4 0.8	2–5	34 20.1%	3.0 1.2	1–6
Kidney *Yang Xu*	14 17.1%	2.3 1.2	1–5	33 19.5%	2.2 1.1	1–5
Kidney *Jing Xu*	2 2.4%	4.5 0.7	4–5	31 18.3%	2.2 1.4	1–6
Spleen Damp	12 14.6%	3.6 2.1	1–7	29 17.2%	3.0 1.5	1–7
Heart Blood *Xu*	5 6.1%	4.2 0.8	3–5	18 10.7%	3.8 1.5	2–8
Liver Wind	5 6.1%	3.2 1.9	1–6	14 8.3%	3.2 1.8	1–6
Liver *Yang* Rising	7 8.5%	4.4 2.6	1–8	11 6.5%	3.5 1.1	2–5
Heart *Qi Xu*	10 12.2%	3.7 1.2	2–5	6 3.6%	2.7 0.5	2–3
Damp Heat Retention	5 6.1%	3.2 1.8	1–5	4 2.4%	3.5 1.3	2–5
Wind—Cold Invasion	0 0.0%	0.0 0.0	0	1 0.6%	3.0 0.0	3–3
Heart *Yang Xu*	3 3.7%	3.3 0.6	3–4	0 0.0%	0.0 0.0	0–0

TCM, Traditional Chinese Medicine; TMD, temporomandibular joint disorder; StDev, standard deviation.

Table 2. Intra-rater Correlation and Fleiss' Kappa by Diagnosis

	Calibration Kappa			*Recalibration Tucson Kappa Overall*	*Recalibration Portland Kappa Overall*
	Wave 1	*Wave 2*	*Wave 3*		
Fleiss' kappa overall agreement	0.112 ± 0.208	0.149 ± 0.206	0.318 ± 0.208	0.618 ± 0.103	0.576 ± 0.189
Whole systems Fleiss' kappa overall agreement	0.231 ± 0.224	0.276 ± 0.218	0.606 ± 0.224	0.688 ± 0.98	0.602 ± 0.161
Inter-rater correlation					
Liver *Qi* Stagnation	0.139 ± 0.220	0.444 ± 0.224	0.565 ± 0.218	1.000 ± 0	1.000 ± 0
Qi and Blood Stagnation	0.681 ± 0.222	0.798 ± 0.224	1.000 ± 0.224	0.258 ± 0.023	0.632 ± 0.343
Liver Blood *Xu*	0.479 ± 0.212	0.028 ± 0.203	0.306 ± 0.203	0.258 ± 0.331	0.632 ± 0.343
Liver Yin *Xu*	0.583 ± 0.224	0.596 ± 0.224	0.615 ± 0.206	0.000 ± 0.354	0.467 ± 0.187
Liver Fire	0.231 ± 0.206	0.692 ± 0.213	0.643 ± 0.209	[a]	[a]

[a]Insufficient cases to calculate.
All figures are mean ± standard error.

Results

In the calibration exercises, all diagnoses were present except Wind-Cold Invasion (Table 1). The frequencies of the diagnoses were similar between the main study and the calibration participants. *Qi* and Blood Stagnation and Liver *Qi* Stagnation were the most severe and most frequent diagnoses when mentioned, whereas Kidney *Jing Xu* and Liver *Yang* Rising were very infrequent but very severe when present.

Using the primary diagnosis only analysis, Fleiss' kappa combined over all three waves of the initial calibration exercise was 0.287 ($p < 0.05$). The kappa for each of the three waves of patients was 0.112, 0.149, and 0.318, showing improvement during the calibration exercise. Landis and Koch[25] have suggested that kappas between 0.10 and 0.20 have slight agreement, 0.21 through 0.40 have fair agreement, 0.41 through 0.60 have moderate agreement, 0.61 through 0.81 have substantial agreement, and 0.81 though 1.00 have almost perfect agreement. It is also in the nature of kappas that fewer categories create higher kappas. The recalibration exercise 18 months later resulted in overall Fleiss' kappa of 0.576 and 0.618 in Tucson and Portland, respectively. The differences between the original exercise and the follow-up were significant at both sites ($p < 0.01$).

Using the kappa of the whole systems approach, agreement was higher over all waves. This may be due to fewer categories. The differences between the two methodologies were not statistically significant but they all trended in the expected direction. The kappa for the initial total calibration exercise was 0.368 (data not shown) and the follow-up exercises were significantly improved at both sites ($p < 0.01$) (Table 2).

Five diagnoses were sufficiently prevalent in the original calibration population over each of the three waves to calculate the inter-rater correlation. The inter-rater correlations show improvement in agreement across all three waves for Liver *Qi* Stagnation and *Qi* and Blood Stagnation, where there were enough diagnoses to calculate the statistic. There was an average agreement of 65%, with the highest agreement in the Liver *Qi* Stagnation diagnosis.

Discussion

In general, this study demonstrated that the inter-rater reliability of TCM diagnoses can be improved through calibration exercises. The original calibration exercise showed better agreement in the second and third waves of participants than the first, after the practitioners gained familiarity with the forms and the diagnostic styles of their colleagues. The agreement further improved at the recalibration exercise, held 18 months later, after the practitioners had considerable experience both with the participants who had TMJD and with the diagnosis and recording process. In all of the exercises, it was much more difficult to achieve agreement in the diagnosis of healthy patients, as their symptom complex is much more subtle.

In the initial kappa analysis and the inter-rater correlations, diagnoses that are related caused disagreement between practitioners. For example, many participants had Liver *Qi* Stagnation and Spleen *Qi* Deficiency or Liver Attacking the Spleen. For these diagnoses, practitioners often disagreed on whether the Liver *Qi* Stagnation or Spleen *Qi* Deficiency was the primary diagnosis. In these cases, agreement could be considered higher than reported as the treatment principle is the same in either case: soothe the Liver and support the Spleen *qi*. In treatment, one has to decide which needs more support and focus the treatment towards bringing balance between these organs. Likewise, the common progression of disease in TCM is from Blood Deficiency to *Yin* Deficiency. A number of practitioners identified patients as being further progressed in the disease cycle, which caused an underreporting of agreement. One can see that a methodology of reporting diagnostic agreement that does not take into account the underlying principles of TCM causes suppressed reporting of agreement among practitioners. In our calibration exercise, counting these as matching diagnoses allowed the overall agreement to rise from 0.287 to 0.368. While the differences were not statistically significant, in all cases the reported agreement rose when taking account of the nature of TCM diagnosis.

In general there was higher agreement for those diagnoses that were more prevalent in the exercise. As Liver *Qi* Stagnation and *Qi* and Blood Stagnation were the two most common diagnoses in the study, practitioners had plenty of practice identifying these cases. Less common diagnoses such as Liver Blood Deficiency and Liver *Yin* Deficiency did not show the same improvements. In future studies, it may be important to balance the calibration patient population by

TCM diagnosis in order to get sufficient practice with each diagnosis.

While the form was tested for face validity in qualitative interviews, the portion of the form used for the practitioners to take notes during the interview process was modified during the original calibration exercise. As practitioners gained more facility with the document, there were places where the form did not match the interview style used by practitioners. Therefore some of the increased agreement may have been from a better questionnaire-guided process. Because the point of the calibration exercise was to solve these problems, this is considered a benefit of the process. The data collection portion of the form never changed throughout the process and thus there was no effect of the form changes on the statistical analysis.

The discussion that happened between waves was an important learning tool for the practitioners and helped them reach consensus on key indicators of each diagnosis while orienting the practitioners towards the scale other practitioners were using. By completing this exercise, practitioners modified the severity score of each diagnosis and agreed upon key determinants of each diagnosis.

Conclusions

Our findings demonstrated that inter-rater reliability of TCM diagnosis can be improved through a training process and a questionnaire-based diagnosis process.[26] Higher levels of improvement may be obtained by eliminating healthy participants, pre-screening calibration participant diagnosis to match the anticipated study population, and providing additional patients for practitioners to evaluate. As with all medical systems, the role of diagnosis within TCM is believed to be vital to the effective application of treatment. In order to evaluate the effectiveness of whole systems of care which include diagnosis and tailoring of treatment to individual needs, calibration exercises and practitioner training should be considered critical to model validity in Oriental medicine and other whole system studies.

Acknowledgments

This publication was made possible by grant number U01AT002570 from the National Center for Complementary and Alternative Medicine (NCCAM) at the National Institutes of Health. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of NCCAM.

Disclosure Statement

No competing financial interests exist.

References

1. Verhoef MJ, Lewith G, Ritenbaugh C, Boon H, Fleishman S, Leis A. Complementary and alternative medicine whole systems research: Beyond identification of inadequacies of the RCT. Complement Ther Med 2005;13:206–212.
2. Verhoef MJ, Vanderheyden LC, Fønnebø V. A whole systems research approach to cancer care: Why do we need it and how do we get started? Integr Cancer Ther 2006;5: 287–292.
3. Ritenbaugh C, Verhoef M, Fleishman S, Boon H, Leis A. Whole systems research: A discipline for studying complementary and alternative medicine. Altern Ther 2003;9: 32–36.
4. Verhoef M, Lewith G, Ritenbaugh C, Thomas K, Boon H, Fønnebø V. Whole systems research: Moving forward. Focus Altern Complemen Ther 2004;9:87–90.
5. Bell I, Koithan M. Models for the study of whole systems. Integrat Cancer Ther 2006;5:293–307.
6. Elder C, Aickin M, Bell I, et al. Methodological challenges in whole systems research. J Altern Complemen Med 2006;12: 843–850.
7. Jonas W, Beckner W, Coulter I. Proposal for an integrated evaluation model for the study of whole systems health care in cancer. Integrat Cancer Ther 2006;5:315–319.
8. Fønnebø V, Grimsgaard S, Walach H, et al. Researching complementary and alternative treatments—the gatekeepers are not at home. BMC Med Res Methodol 2007;7:7.
9. Schaechter J, Neustein SM. P6 acupuncture point stimulation for prevention of postoperative nausea and vomiting. Anesthesiology 2008;109:155–156; author reply, 157–158.
10. Neri I, De Pace V, Venturini P, Facchinetti F. Effects of three different stimulations (acupuncture, moxibustion, acupuncture plus moxibustion) of BL.67 acupoint at small toe on fetal behavior of breech presentation. Am J Chin Med 2007;35:27–33.
11. Neri I, Fazzio M, Menghini S, Volpe A, Facchinetti F. Nonstress test changes during acupuncture plus moxibustion on BL67 point in breech presentation. J Soc Gynecol Investig 2002;9:158–162.
12. Fireman Z, Segal A, Kopelman Y, Sternberg A, Carasso R. Acupuncture treatment for irritable bowel syndrome. A double-blind controlled study. Digestion 2001;64:100–103.
13. Cheng XN. Chinese Acupuncture and Moxibustion. Beijing: Foreign Language Press, 1987.
14. Maciocia G. The Foundations of Chinese Medicine. London: Churchill Livingstone, 1989.
15. Zhang G, Bausell B, Lao L, et al. Assessing the consistency of TCM diagnosis: An integrative approach. Altern Ther Health Med 2003;9:66–71.
16. Sung J, Leung WK, Ching J, Lao L, et al. Agreements among Traditional Chinese Medicine practitioners in the diagnosis and treatment of irritable bowel syndrome. Aliment Pharmacol Therapeutics 2004;20:1205–1210.
17. Kim M, Cobbin D, Zaslawski C. Traditional Chinese medicine tongue inspection: An examination of the inter- and intrapractitioner reliability for specific tongue characteristics. J Altern Complement Med 2008;14:527–536.
18. Baker J, Ben-Tovim DI, Butcher A, Esterman A, McLaughlin K. Development of a modified diagnostic classification system for voice disorders with inter-rater reliability study. Logoped Phoniatr Vocol 2007;32:99–112.
19. Weyer A, Abele M, Schmitz-Hübsch T, et al. Reliability and validity of the scale for the assessment and rating of ataxia: a study in 64 ataxia patients. Mov Disord 2007;22:1633–1637.
20. Gur AY, Lampl Y, Gross B, Royter V, Shopin L, Bornstein NM. A new scale for assessing patients with vertebrobasilar stroke-the Israeli Vertebrobasilar Stroke Scale (IVBSS): Inter-rater reliability and concurrent validity. Clin Neurol Neurosurg 2007;109:317–322. Epub 2007 Jan 24.
21. Berk M, Malhi GS, Cahill C, et al. The Bipolar Depression Rating Scale (BDRS): Its development, validation and utility. Bipolar Disord 2007;9:571–579.
22. Rösler M, Retz W, Retz-Junginger P, et al. Attention deficit hyperactivity disorder in adults. Benchmarking diagnosis

using the Wender-Reimherr adult rating scale Nervenarzt 2008;79:320–327.
23. Ritenbaugh C, Hammerschlag R, Calabrese C, et al. A pilot whole systems clinical trial of Traditional Chinese Medicine and naturopathic medicine for the treatment of temporomandibular disorders. J Altern Complement Med 2008;14:475–487.
24. Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bul 1971;76:378–382.
25. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–174.
26. Zhang GG, Singh B, Lee W, Handwerger B, Lao L, Berman B. Improvement of agreement in TCM diagnosis among TCM practitioners for persons with the conventional diagnosis of rheumatoid arthritis: Effect of training. J Altern Complement Med 2008;14:381–386.

Address correspondence to:
Scott Mist, Ph.D., L.Ac.
Oregon College of Oriental Medicine
1924 NE 56th Avenue
Portland, OR 97213

E-mail: scott.mist@comcast.net