Chance level performance in expert diagnoses with applied kinesiology

Jennifer R. McCullen[a,c,*], **Beth Baribault**[b,c,*], **and Joachim Vandekerckhove**[c,*,†]

Draft of October 15, 2025.

We test the diagnostic validity of manual muscle testing (MMT). The MMT is an alternative-medicine technique used to assess the suitability of natural remedies for minor medical ailments. The technique involves a trained tester gauging a patient's muscle resistance while the patient holds a candidate supplement close to the body, and supplements are then recommended on the basis of between-trial changes in muscle resistance. MMT is widely used despite the lack of a known underlying physical mechanism. In a pre-registered study, we evaluate the ability of practitioners to reliably rank supplements in order of suitability (i.e., in terms of their positive effect on perceived muscle resistance). We provide details of a custom analysis in which we quantify the evidence for competing accounts. The data are overwhelmingly more consistent with the supposition that the rankings are random than with any competing account.

applied kinesiology, alternative medicine, muscle testing, pseudoscience, open science

Manual muscle testing (MMT) is a diagnostic technique used in alternative medicine to assess the suitability of natural food supplements (e.g., tea, bee pollen. . .) for minor medical conditions (e.g., chronic fatigue). Several schools of applied kinesiology (Eden, 2008) involve a belief that *mere proximity* of such remedial supplements has a measurable effect on a patient's muscle tone, so that the suitability of a remedy can be predicted by the effect it has on patient muscle strength. According to a recent survey (Jensen, 2015), some 200,000 practitioners adhere to one of these schools.

The manual muscle test. The use of MMT as a diagnostic procedure involves exposing the patient to various substances in order to detect either a remedy or toxin. A weakening of muscle tone is interpreted as a signal that the substance is a toxin to the patient whereas a tonifying (i.e., strengthening) response is taken to signify a potential remedy. Substances can be variously tested through ingestion, insalivation, or non-local proximity (NLP; Schwartz et al., 2014). NLP testing relies on the assumption that the mere presence of a substance can cause a change in muscle strength. Procedures that rely on NLP have the advantage of avoiding the need for ingestion or topical application (i.e., they are entirely non-invasive). NLP testing is commonly performed using a standard "arm pull down" muscle test (see Fig. 1). For a broader review of the MMT and applied kinesiology that goes beyond NLP, see Haas, Cooperstein, and Peterson (2007).

To perform a standard muscle test using the arm-pull-down technique, a patient is asked to stand and hold a substance (a potential remedy or toxin) next to the body with one arm, while extending the opposite arm, palm down. The muscle tester will then apply pressure to the back of the extended hand and judge whether the muscle response was "strong" or "weak" relative to trials in which other substances are held.

Fig. 1. A muscle testing trial. The participant (right) is standing with one arm outstretched and holding a vial while the muscle tester (left) applies gentle downward pressure on the outstretched arm. From Schwartz et al. (2014), used with permission.

Previous research. While there is some evidence to support the notion that experts are able to detect trial-to-trial changes in muscle tone (Florence et al., 1984; Pollard, Lakay, Tucker, Watson, & Bablis, 2005; Schmitt & Leisman, 1998), previous studies on the efficacy of the non-local manual muscle test have yielded conflicting results, with some supporting the validity of the test (Radin, 1984) and others finding inconclusive results (Arnett, Friedenberg, & Kendler, 1999; Keating, Kendler, & Merriman, 2004; Lüdtke, Kunz, Seeber, & Ring, 2001; Quintanar & Hill, 1988).

Unfortunately, while Radin's finding of a nonlocal effect has failed to replicate, these failures to replicate are weakened by their reliance on classical null hypothesis significance testing (NHST). NHST allows researchers to reject, but never confirm, null hypotheses: The procedure starts with the assumption that there is no effect (the null hypothesis) and then uses that assumption to compute the probability p that data as extreme as, or more extreme than, the real data would be observed upon repeated execution of the exact same study. But, since the null hypothesis is *assumed*, it cannot be *concluded*. Here, we will use a comparative procedure in which we use Bayes factors to quantify evidence for (or against) the null hypothesis.

[a]University of Puget Sound; [b]University of California, Berkeley; [c]University of California, Irvine
All authors contributed to the final draft.
[*]All authors contributed equally to this work.
[†]To whom correspondence should be addressed. E-mail: joachim@uci.edu.
The authors declare no conflicts of interest. The authors were supported by National Science Foundation grants #1230118 and #1534472 (JV and BB). The authors thank Emily Young and Hannah Guild for their help with data collection.

Methods

Participants. We recruited two groups of participants: "patient" participants who reported suffering from fatigue, and "muscle tester" practitioners trained in MMT techniques. A flowchart of participant recruitment and inclusion is shown in Figure 2. Patient participants responded to fliers posted in the Orange County community. Of 38 individuals who were screened online, 29 met the criterion for fatigue based on the Fatigue Assessment Scale (Michielsen, De Vries, & Van Heck, 2003). Of these, 18 (10 female, 8 male; mean age 24.3) consented to participate in the testing phase. Muscle testers were recruited via fliers and direct contact. Nine practitioners initially participated; however, the data from two were excluded from analysis because they were uncomfortable with the 'arm pull down' technique as described and did not follow the ranking protocol, leaving a final sample of seven practitioners (4 male, 3 female; ages 44 to 72). The typical muscle tester in the final sample had 10 years of experience (range: 2 to 37 years).

Materials. For the to-be-tested substances, we selected an assortment of five herbal remedies commonly used to treat fatigue: ashwagandha, bee pollen, yerba mate, eleuthero, and green tea. These herbs are commonly used in an attempt to increase energy and were selected because of their placement in adaptogenic energy sections in local health food stores in Orange County, CA. Two grams of each herb and two cotton balls were placed into opaque bottles and sealed. These bottles were randomized and labeled A through E. Indistinguishability of the different bottles was confirmed by the senior author (JV) who was not present during the assembly of the bottles.

"Patient" participants were provided a description of the muscle testing procedure and asked whether they believed that muscle testing has the ability to determine if a supplement is needed by the body. The description and question are available via the Open Science Framework (OSF; `osf.io/qe9p4`).

Procedure. Testing was divided into two separate two-hour sessions. In the first session five muscle testers performed muscle tests on each of ten patients. In the second session, four muscle testers performed tests on each of eight patients. Prior to testing, patient and muscle testers were led into separate rooms to discuss the procedure. Patients were debriefed on how the test is done. Muscle testers were briefed on the technique being used, the materials to use during testing, and how to record results.[1] At this point, two muscle testers (both in the second session) were no longer comfortable with the "arm pull down" technique as described in the recruitment email. Their data were incomplete at the end of the session and were not used for the analyses. When testing began, all muscle testers were in separate rooms and participants moved between rooms to interact with each muscle tester.

Each muscle tester received a set of five herbs in opaque bottles, randomly labeled A-E. The testers were asked to test each patient in their session using these five bottles and then rank the bottles in order of their suitability for that patient. After testing a patient, the testers were asked to indicate with a simple Yes/No question whether they were confident in their judgment. The response sheet is available on OSF (`osf.io/ysv92`).

Once the testers finished testing every patient in their session, the bottles were re-randomized (i.e., the testers were given a different set of bottles with the same substances but a different random set of labels A-E) and the testers re-tested all patients.

[1] During the first testing session, some practitioners expressed initial hesitancy about using opaque bottles, but after the first round of testing, they reported confidence in their ability to perform the muscle test under the experimental conditions.

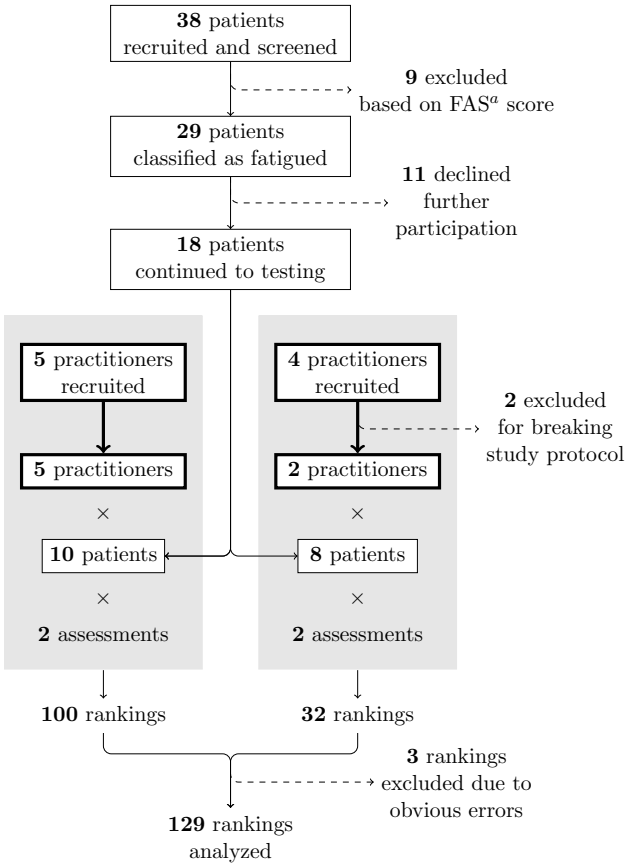

Fig. 2. Data collection flowchart. Each shaded block represents a day of in-person testing where each patient participant (thin outline) was assessed twice by each applied kinesiology practitioner (thick outline). All exclusions are explained in-text. [a]: FAS = Fatigue Assessment Scale.

The experimental protocol was preregistered on OSF (`osf.io/wymjr`) and approved by the Institutional Review Board of the University of California, Irvine (HS# 2015-1926).

Model-based data analysis. In this section, we provide full detail regarding our custom model-based data analysis. This section concludes with an informal conceptual summary that avoids most of the technical detail.

To avoid issues with null hypothesis significance testing (i.e., the inability to confirm null hypotheses; Wasserstein & Lazar, 2016), we opted for a model-comparison approach to analyze the consistency of muscle test results. If the muscle testing procedure is sensitive to a true signal, rank orders given by two muscle testers for the same patient should agree better than chance. Similarly, rank orders provided by a single muscle tester on two occasions should agree better than chance.

In order to quantify agreement between and within muscle testers, we used Kendall's (1938) scoring rule for ordered lists. One version of the rule involves determining the score Kendall's τ, which is the total number of adjacent pairwise swaps required to move from a given order to a target order.[2]

A family of distributions for Kendall's τ In the absence of biases, and assuming the complete absence of agreement, all rank orderings of k items—of which there are $k!$ possible—are

[2] This is one of a few current definitions of τ. Other varietals are linearly transformations of this one.

Table 1. Probability distributions of Kendall's τ, by sequence length k, assuming no agreement. Each column gives the probabilities for different values of τ assuming a sequence of length k. The bottom row gives $n!$, the total number of possible orderings for a list of length n. This distribution is based on the so-called Mahonian triangle (MacMahon, 1915).

	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
$\tau=0$	1	1/2	1/6	1/24	1/120
$\tau=1$	.	1/2	2/6	3/24	4/120
$\tau=2$	.	.	2/6	5/24	9/120
$\tau=3$	.	.	1/6	6/24	15/120
$\tau=4$	.	.	.	5/24	20/120
$\tau=5$	.	.	.	3/24	22/120
$\tau=6$	.	.	.	1/24	20/120
$\tau=7$	.	.	.	.	15/120
$\tau=8$	.	.	.	.	9/120
$\tau=9$	.	.	.	.	4/120
$\tau=10$	.	.	.	.	1/120
$k!$	1	2	6	24	120

equally likely. This implies an expected probability distribution for τ. The expected probabilities of all scores τ, by list length, are given in Table 1. The probability of observing a given τ with a certain list length k can now be determined via this lookup table. We denote the probability as $F_k(\tau)$, so that $F_4(3) = 6/24 = 0.25$ (because 6 of the 24 possible orderings of a four-item sequence have a Kendall's τ of 3).

These distributions of τ in the absence of agreement and bias can serve to construct a new family of distributions of τ. First, the multinomial distribution $F_k(\tau)$ expresses the probability of observing a given value of τ in a list of length k: $P(\tau \mid k) = F_k(\tau)$.

In order to perform model comparison, it will be useful to define alternative distributions of τ in which the assumption of no agreement is relaxed. We define a *noncentral* version of the $F_k(\tau)$ distribution family whose noncentrality parameter ν indicates *the number items on whose rank order a pair of raters agree* (a pair of muscle testers may agree on the position of ν items). ν is a non-negative integer and $\nu < k$. The distributions satisfy

$$F_k(\tau \mid \nu) = F_{k-\nu}(\tau \mid 0) = F_{k-\nu}(\tau)\,, \qquad [1]$$

meaning that for each agreed-upon item, the probability distribution of ν shifts one column to the left (in Table 1).

If $\nu = 0$, there is no agreement and the expected distribution reduces to the central $F_k(\tau)$ distribution. If $\nu = k-1$, there is complete agreement—because the n^{th} item is then determined—and τ must be 0.

Inferring the agreement ν The new parameter ν has a straightforward interpretation as the number of extreme items on which a pair of participants agree. Given a set of k observed scores $T_k = (\tau_1, \tau_2, \ldots, \tau_k)$, the joint likelihood is

$$F_k(T_k \mid \nu) = \prod_{i=1}^{k} F_k(\tau_i \mid \nu)$$

and with a conjugate multinomial prior distribution for ν, the posterior distribution of ν assuming the model $F_k(\cdot)$ is

$$P(\nu \mid F_k(\cdot)\,, T_k) = \frac{P_\nu(\nu) \prod_{i=1}^{k} F_k(\tau_i \mid \nu)}{\sum_{s=0}^{k-1} P_\nu(s) \prod_{i=1}^{k} F_k(\tau_i \mid s)}. \qquad [2]$$

A lapse rate λ The multinomial likelihood function $F_k(\cdot \mid \nu)$ has the property that it is zero for values $\tau > \frac{1}{2}(k-\nu)\,(k-\nu-1)$. While it is true that these values are impossible if participants agree on ν or more items and make no reporting errors, we can allow for reporting errors by introducing a *lapse rate* λ, leading to the following likelihood:

$$G_k(T_k \mid \nu, \lambda) = \prod_{i=1}^{k} \left[(1-\lambda)\, F_{k-\nu}(\tau_k) + \lambda F_k(\tau_k)\right]. \qquad [3]$$

That is, with probability $(1-\lambda)$ the response comes from the regular process, so the likelihood is $F_{k-\nu}(\tau)$, but with probability λ the response comes from the completely random process, so the likelihood is $F_k(\tau)$. The lapse rate λ could be estimated, integrated away as a nuisance variable, or can be set to a convenient small value such as .05. We choose to treat λ as a nuisance variable and assign it a uniform distribution from 0 to $\frac{1}{2}$; that is, we do not believe that more than 50% of data points will be the result of a lapse. The density is then $p_\lambda(\cdot) = 2$, and the nuisance variable is treated by taking a weighted average of the likelihood in Equation 3 using this prior as weight. Since λ is continuous, the weighted average is an integral over λ:

$$G_k(T_k \mid \nu) = \int_0^{\frac{1}{2}} p_\lambda(\lambda) \prod_{i=1}^{k} \left[(1-\lambda)\, F_{k-\nu}(\tau_k) + \lambda F_k(\tau_k)\right] d\lambda.$$

Combining this marginalized likelihood with the prior for ν and normalizing yields the posterior probabilities of ultimate interest:

$$P(\nu \mid G_k(\cdot)\,, T_k) = \frac{P(\nu) G_k(T_k \mid \nu, \lambda)}{\sum_{s=0}^{k-1} P(s) G_k(T_k \mid \nu, \lambda)}. \qquad [4]$$

Prior probability and Bayes factors In order to obtain a posterior probability (as in Eq. 4), we need to define an a-priori probability for each value of ν. These prior probabilities simply indicate the strength of our belief in each possible ν-value—but they are necessarily subjective. We address this subjective element in two ways.

First, we defined three reasonable prior distributions: a *flat prior* under which each value for ν is equally likely (20%); a *skeptical prior* under which the model of no correspondence ($\nu = 0$) is most likely (50%) and the other four models are equally likely (12.5%); and an *adherent prior* under which the model with complete correspondence ($\nu = 4$) is most likely (50%) and the four other models are equally likely (12.5%). To assess the influence of prior beliefs in ν, we compare the outcome of our analysis under each scenario. It will turn out that our conclusions are robust to the choice of prior.

Second, we will compute a *Bayes factor* (Kass & Raftery, 1995). While our ultimate interest is in the posterior probability or the *posterior ratio*—the relative probability that $\nu = 0$ versus $\nu > 0$, where the latter's probability is the sum of the posterior probabilities for all nonzero values of ν—this ratio can be written as a product of the *prior ratio* and some factor B_{01}:

$$\frac{P(\nu = 0 \mid G_k, T_k)}{P(\nu > 0 \mid G_k, T_k)} = \frac{P(\nu = 0)}{P(\nu > 0)} \times B_{01} \qquad [5]$$

In Equation 5, B_{01} is known as the *Bayes factor* and will indicate the degree to which the data support the "random-assignment hypothesis" ($\nu = 0$) over the alternative ($\nu > 0$). Importantly, B_{01} is independent of the prior probabilities of ν and is therefore an

attractive quantification of evidence in the data. The Bayes factor is also symmetric, in that it can simply be inverted to obtain the degree of support for the alternative hypothesis ($B_{10} = B_{01}^{-1}$). Conventionally, a Bayes factor of 10 or more indicates strong evidence.

Conceptual summary of the analysis technique We developed a novel procedure to quantify the consistency in muscle test results both across muscle testers and across testing occasions. We consider two broad scenarios: a "random-response model" under which rank orders agree only by chance and an "alternative model" under which rank orders agree more than would be expected by chance. We will calculate a *Bayes factor* that expresses the relative evidence for the random-response model over the alternative model. A Bayes factor of 10 or more indicates strong evidence.

Results

The Bayes factor favoring the random-response model is approximately 1.6 trillion ($B_{01} \approx 1.620 \times 10^{12}$), indicating overwhelmingly strong evidence. Accordingly, the posterior probability of no agreement ($\nu = 0$) approaches 1 for all prior distributions we consider.

For a graphical illustration of this model selection result, Figure 3 depicts with lines the expected probability distribution of τ for all possible values of ν. The distributions widen and flatten as agreement ν goes to 0, and they peak at complete agreement ($\tau = 0$) as ν goes to $k - 1$. The bars indicate the observed distribution in our sample. Visual inspection reveals a striking correspondence between the data and the random-response model of no correspondence.

For this analysis, we excluded trial pairs in which the tester did not indicate strong confidence in their judgment (10 pairs) and trial pairs in which the patient indicated that they did not believe in the effectiveness of muscle testing (42 pairs). This left a total of 83 trial pairs (out of 129 initially). All changes to our censoring rule (e.g., also including trials with low tester confidence or trials with skeptical patients) led to Bayes factors that *more strongly* supported the random-response model (with $B_{01} \approx 2.152 \times 10^{21}$ if no censoring is performed at all), so we do not elaborate on these decisions.

A similar test can be applied to the within-tester consistency (between the two occasions on which the same patient was tested by the same tester). For this scenario, $B_{01} \approx 2.472 \times 10^{3}$ with censoring and $B_{01} \approx 5.782 \times 10^{6}$ without.

Sample analysis. Given that approximately 200,000 practitioners reportedly use MMT techniques, one might question whether our sample of 7 practitioners was representative.[3] If a substantial proportion of practitioners could in fact perform the MMT, the probability of sampling 7 practitioners who could not would be extremely low. For example, even if only 50% of practitioners could do the task, the probability of sampling 7 who cannot would be less than 1%. Our finding is far more consistent with MMT being ineffective with all practitioners.

Discussion

The practice of manual muscle testing as a diagnostic tool implies a perceptual challenge for the practitioner. The reliability of the test can be evaluated via the correspondence between repeated tests (the same tester with the same patient) and via the correspondence between testers (with the same patient). On both measures, the degree of correspondence was overwhelmingly more consistent with a random-response model than with any model that assumed the detection of a signal. The assignment of supplements by our trained testers was random despite testers' reported confidence in the outcomes and participants' reported faith in the procedure. Our experiment delivers strong evidence that manual muscle testing with non-local proximity fails as a diagnostic procedure.

This work has certain implications for testing controversial claims. Our experiment was designed not only to test a specific claim but also to serve as a case study for fair evaluation of controversial topics. From this experience, we offer three recommendations. First, consulting with proponents of a claim during the design phase can preemptively address methodological criticisms. (Our decision to include confidence ratings resulted from these consultations.) Second, statistical methods capable of providing evidence for the null hypothesis are critical; conventional significance testing would have only allowed us to 'fail to reject' the null, a much weaker conclusion. Finally, maintaining transparency through preregistration and open data can build trust in the final result, regardless of the outcome.

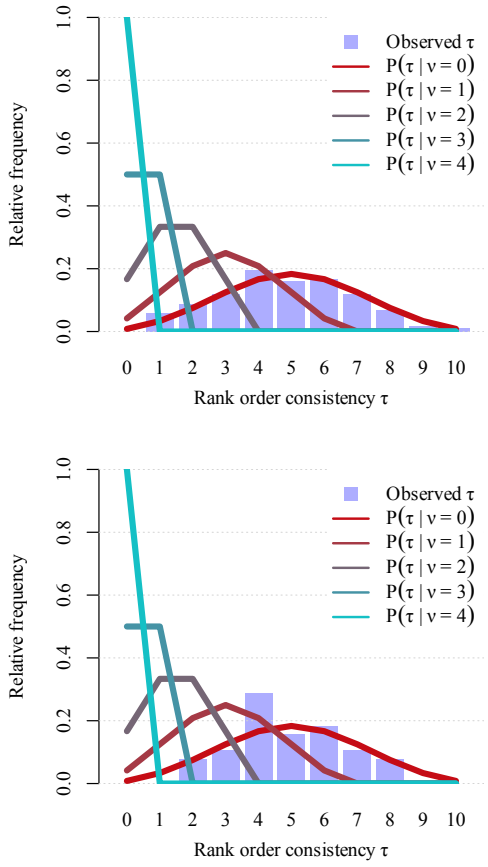

Fig. 3. Distributions of τ as predicted by each of the five models under consideration (lines), overlaying the observed distribution of τ in the experiment (bars). The lines are graphical representations of the model-predicted expected probabilities, which were computed by combining the probabilities in Table 1 with a lapse process that's given by the rightmost column in that table. The data appear to correspond most closely to the model assuming no agreement ($\nu = 0$) both for the global correspondence analysis (top) and the internal consistency analysis (bottom).

[3] Given the large Bayes factor, we do not have concerns about statistical power.

Open science statement. The raw data from this study, along with custom R code used for the statistical analysis, are freely available via `https://codeocean.com/capsule/0458352/tree/v1`.

References

Arnett, M. G., Friedenberg, J., & Kendler, B. S. (1999). Double-blind study of possible proximity effect of sucrose on skeletal muscle strength. *Perceptual & Motor Skills*, *89*, 966–968.

Eden, D. (2008). *Energy medicine: Balancing your body's energies for optimal health*. New York, NY: Penguin Group.

Florence, J. M., Pandya, S., King, W. M., Robison, J. D., Signore, L. C., Wentzell, M., & Province, M. A. (1984). Clinical trials in Duchenne dystrophy: Standardization and reliability of evaluation procedures. *Physical Therapy*, *64*(1), 41–45.

Haas, M., Cooperstein, R., & Peterson, D. (2007). Disentangling manual muscle testing and applied kinesiology: Critique and reinterpretation of a literature review. *Chiropractic & Osteopathy*, *15*(1), 11.

Jensen, A. M. (2015). Estimating the prevalence of use of kinesiology-style manual muscle testing: A survey of educators. *Advances in Integrative Medicine*, *2*(2), 96–102.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

Keating, T. M., Kendler, B. S., & Merriman, W. (2004). Evaluation of a possible proximity effect of aspartame and vitamin C on muscular strength. *Perceptual & Motor Skills*, *98*, 100–102.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, *30*(1-2), 81-93.

Lüdtke, R., Kunz, B., Seeber, N., & Ring, J. (2001). Test-retest-reliability and validity of the kinesiology muscle test. *Complementary Therapies in Medicine*, *9*(3), 141–145.

MacMahon, P. (1915). *Combinatory Analysis, Volumes I and II*. AMS Chelsea Pub.

Michielsen, H. J., De Vries, J., & Van Heck, G. L. (2003). Psychometric qualities of a brief self-rated fatigue measure: The Fatigue Assessment Scale. *Journal of Psychosomatic Research*, *54*(4), 345–352.

Pollard, H., Lakay, B., Tucker, F., Watson, B., & Bablis, P. (2005). Interexaminer reliability of the deltoid and psoas muscle test. *Journal of Manipulative & Physiological Therapeutics*, *28*, 52–56.

Quintanar, A. F., & Hill, T. V. (1988). Sugar proximity and human grip strength. *Perceptual and Motor Skills*, *67*(3), 853–854.

Radin, D. I. (1984). A possible proximity effect on human grip strength. *Perceptual and Motor Skills*, *58*(3), 887–888.

Schmitt, W. H., & Leisman, G. (1998). Correlation of Applied Kinesiology muscle testing findings with serum immunologobulin levels for food allergies. *International Journal of Neuroscience*, *96*(3-4), 237–244.

Schwartz, S. A., Utts, J., Spottiswoode, S. J. P., Shade, C. W., Tully, L., Morris, W. F., & Nachman, G. (2014). A double-blind, randomized study to assess the validity of Applied Kinesiology (AK) as a diagnostic tool and as a nonlocal proximity effect. *Explore*, *10*(2), 99–108.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician*, *70*(2), 129-133.