Inter-Examiner Reliability Studies in an Educational Setting: How & Why to D.O. Them

Professor Michael L. Kuchera, DO, FAAO
Dept of Osteopathic Manipulative Medicine
Philadelphia College of Osteopathic Medicine
michaelkuc@pcom.edu

Goals of This Workshop

1. Osteopathic educational departments/institutions should recognize the **value** in prioritizing & performing reproducibility studies (RS).
2. Knowledge of the **phases** used in RS.
3. Recognition of the **pitfalls** in conducting RS.
4. Overview of the best **statistical method (kappa)** for interexaminer reliability (RS)
5. Appreciate the use of the **FIMM Protocol** to avoid the "prevalence pitfall"
6. Perform a **mock RS** for a team-selected diagnostic test
7. To increase the **evidence base** for osteopathic (and other manual medicine) diagnostic tests.

Value of Reproducibility Studies in Osteopathic & Manual Medicine

- Fédération Internationale de Médecine Manuelle **(FIMM)** recommends Reproducibility Studies as the **#1 research priority** for National M/M Societies
- If procedures used to **identify/diagnose somatic dysfunction** are not tested, how can we test the efficacy of methods to treat somatic dysfunction?

Recent Full Course Sponsored by the BSO, BIMM & the IAMMM

"TEST THE TEST"
Practical interdisciplinary course in reproducibility studies

Sunday 10th April 2011, 9 a.m.—17 p.m.
British School of Osteopathy
275 Borough High Street, London, SE1 1JE.

SUPPORTED BY
BIMM
THE BRITISH SCHOOL OF OSTEOPATHY

Next IAMMM Course ...

Special Value in Prioritizing & Conducting Reliability Tests in Osteopathic Educational Settings

Personally, conducting these tests have made me:

- ... a better **teacher**.
- ... a better **researcher**.
- ... a more **attentive learner**.

- Describing/Demonstrating Procedures
- Identifying Critical Performance Steps
- Watching Students for Key Mis-Steps
- Communicating a Rational, "Why"
- Standardizing How Taught in Dept

Special Value in Prioritizing & Conducting Reliability Tests in Osteopathic Educational Settings

Faculty Research Potential

- **Meaningful Research**: High priority for profession
- **Publishable**: Desire to publish quality studies
- **Inexpensive** to conduct
- **Reproducible** process
- Available & willing **subject pool** (caveats)
- All of the above for endless **student research** projects

Student Benefits

- Involves students in research early on
- Demonstrates faculty commitment to research & education
- Enhances student respect for attention to learning detail for hands-on testing
- Better understanding of expectations

Nomenclature:
Reliability α Reproducibility & Validity

Reproducibility reflects the extent of agreement between examiners using the same test on the same subject (inter-examiner) or the use of the same test by the same examiner at different times (intra-examiner).

Validity measures the extent to which a diagnostic test actually tests what if is supposed to test. (How well does it stand up to a "gold standard?")

Reproducibility

Validity

Reporting & Analyzing Findings

- **Nominal Data**
  - Yes – No
  - Kappa Value Best
- **Ordinal Data**
  - Normal – Slight – Marked
  - SD Severity 0-1-2-3
  - Weighted Kappa Best
- **Interval or Continuous Data**
  - Report Degrees of Restriction for Example
  - Use Student T-Test or ANOVA

Best Statistical Analysis for InterExaminer Reliability (Reproducibility) Testing is to Gather Nominal Data for Use in Calculating Kappa. Think How to Phrase Questions Asked About Test or Group of Tests Leading to Single "Yes-No" or Single "Right-Left" (etc)

Primary Resource for Reliability Tests

Document Basic Diagnostic Aspects

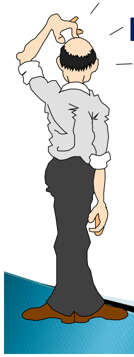
FIMM

FIMM INTERNATIONAL ACADEMY OF MANUAL/MUSCULOSKELETAL MEDICINE

FIMM Scientific Committee: 12 Golden Rules for Manual Medicine Research & Protocol for Inter-Examiner Reliability (Kappa)

www.fimm-online.com ← #1 priority

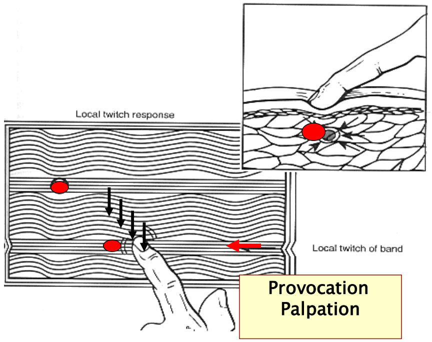
Defining Inter-Examiner Reliability in Palpation



Kappa Caveat: *"#@*&! statistics"*

- In assessing kappa (Κ)
  - Ideal "test" population:
    - 50% with – 50% without characteristic
  - Kappa of 0.50
    - Midway chance & complete agreement
  - Uneven split for testing?
    - Poor (low) kappa regardless
    - FIMM Protocol corrects (n=40; 2 examiners)

Examination Myofascial Trigger Points



Provocation Palpation

Local twitch of taut band when stimulate the local spot with provocation palpation (perpendicular to fiber direction)

Use dolorimeter (or algometer) for standard pressure to elicit pain

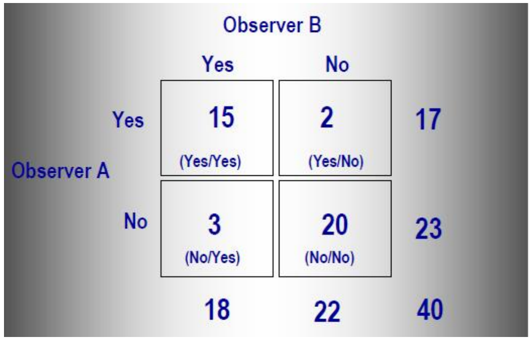
Reproducible with Good Kappa Values Differ by Point Tenderness best Κ
Simons & Mense

2X2 Contingency Table

		Observer B		
		Yes	No	
Observer A	Yes	a (Yes/Yes)	b (Yes/No)	a+b
	No	c (No/Yes)	d (No/No)	c+d
		a+c	b+d	n

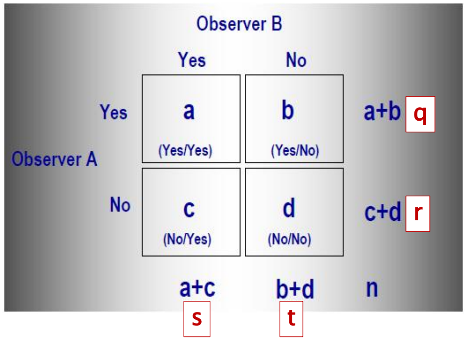
- Entering the data is easy
- Take list of subjects with data from Observers A&B and enter into table

Example



		Observer B		
		Yes	No	
Observer A	Yes	15 (Yes/Yes)	2 (Yes/No)	17
	No	3 (No/Yes)	20 (No/No)	23
		18	22	40

2X2 Contingency Table

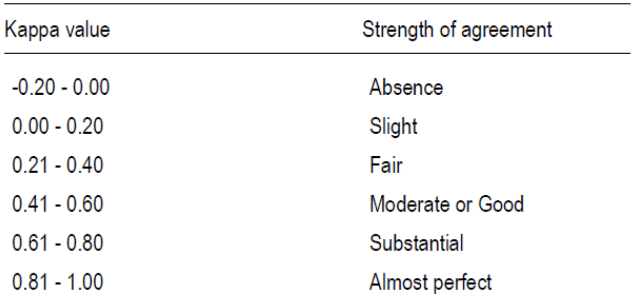


		Observer B		
		Yes	No	
Observer A	Yes	a (Yes/Yes)	b (Yes/No)	a+b q
	No	c (No/Yes)	d (No/No)	c+d r
		a+c s	b+d t	n

- P_o (Observed Agreement) = a +d (yes:yes) + (no:no)
- P_e (Expected by Chance Agreement) = ([a+c]*[a+b]) + ([b+d]*[c+d]) ...sq+tr
- Kappa =

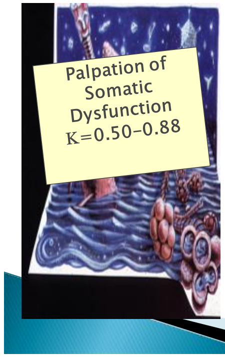
$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

Landis & Koch: Interpreting Inter-Examiner Reliability Statistics



Kappa value	Strength of agreement
-0.20 - 0.00	Absence
0.00 - 0.20	Slight
0.21 - 0.40	Fair
0.41 - 0.60	Moderate or Good
0.61 - 0.80	Substantial
0.81 - 1.00	Almost perfect

Example: Reliability of Routine Physical Examination Tests by Physicians in the Pulmonary System



Palpation of Somatic Dysfunction K=0.50-0.88

Physical Examination: Respiratory System

- Wheezes K=0.51
- Crackles K=0.41
- Bronchial breathing K=0.32

Percussion (CXR gold standard)

- Texts agree not sensitive >5cm below chest wall or <3cm in size
- Sensitivity = 15.4%
- Specificity = 97.3%
- Percussion K=0.50

Somatic Dysfunction: Researching Palpation & Kappa

Must Have Training / Consensus Standardization

- Kappa > 0.40 sought
- Palpatory Diagnostics (Lumbar)
  - 0.88 Tenderness*
  - 0.72 Asymmetry–Segmental rotation
  - 0.50 Restricted motion–Segmental rotation
  - 0.55 Tissue Texture Change*

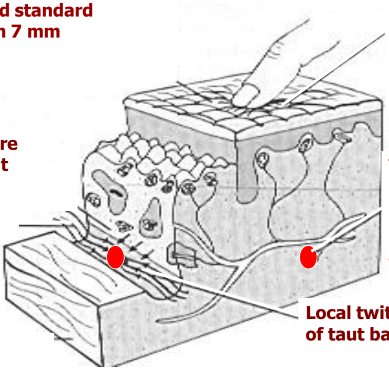
Degenhardt et al: *JAOA 102(8):* 439, Aug 2002

Sciotti: *Pain 93:259-66,* 2001

Reproducibility: Trapezius MTrP

Criterion reliability threshold>80% agreement

- 4 blinded **experts** + pretest
- Algometry gold standard
- Located within 7 mm



Spot Tenderness over TrP
κ=0.95

Tissue texture change (Taut band)
κ=0.98

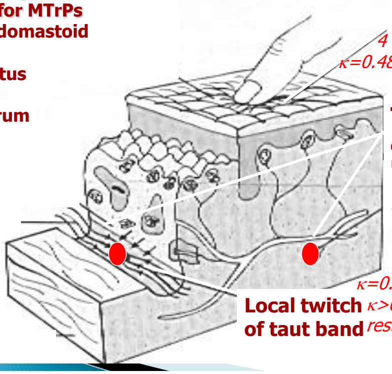
Tissue texture change (Palpable nodule)
κ=0.99

Local twitch of taut band
κ poor X

Gerwin: *Pain 69:65-73,* 1997

Palpation of Myofascial Dysfunction

- Criterion reliability threshold>80% agreement
- 5 blinded experts + pretest
- Muscles: differing reliable criteria
- 5 Paired Sites for MTrPs
  - Sternocleidomastoid
  - Trapezius
  - Infraspinatus
  - Latissimus
  - Ext Digitorum



Spot Tenderness
4 ms κ>0.60
κ=0.48 infraspin

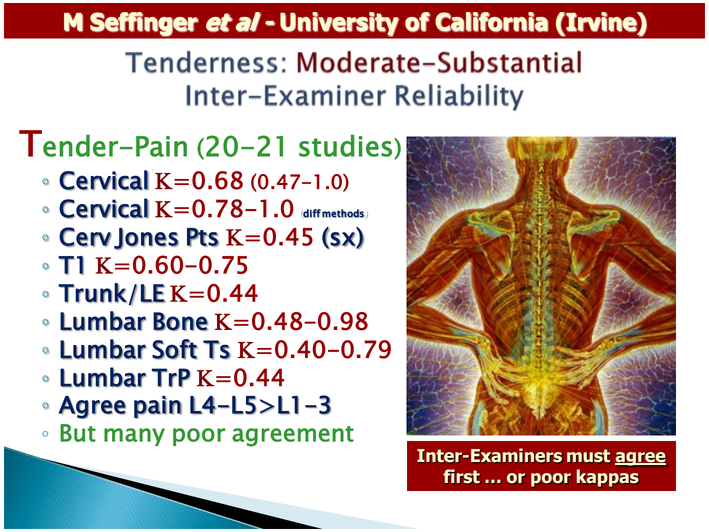
Tissue texture change (Taut band)
κ=0.40 (2 ms)
rest κ>0.60

Local twitch of taut band
κ=0.57 Lat dors
κ>0.80 Ext dig
rest poor reliab

Annotated Bibliography: Inter-Examiner Reliability

OH, OH! NOW YOU'RE GONNA GET IT! YOU KNOW YOU'RE SUPPOSED TO DO IT ON THE PAPER!
©1987 MARK

- Literature: Content Validity & Reliability 1966–2002
  - Content Validity 5 articles; Reliability 59 articles
- Reliability grouped:
  - T: Pain provocation tests
  - A: Anatomic landmarks
  - R: Motion tests
  - T: Paraspinal soft tissue tests



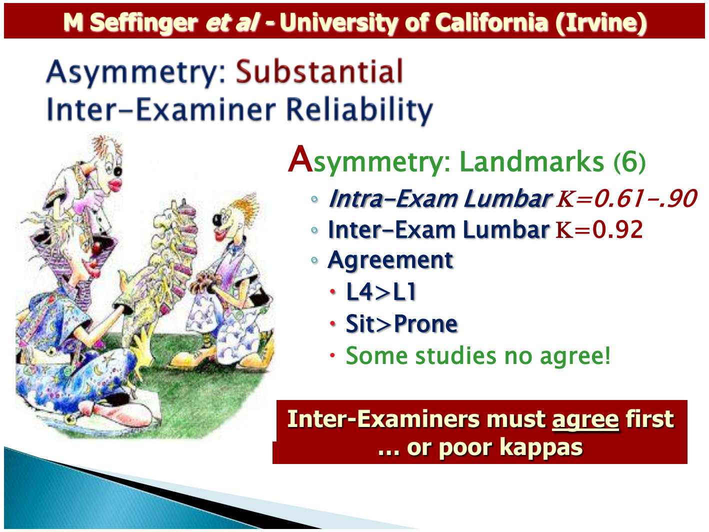
M Seffinger *et al* - University of California (Irvine)

Tenderness: Moderate-Substantial Inter-Examiner Reliability

Tender-Pain (20-21 studies)

- Cervical K=0.68 (0.47-1.0)
- Cervical K=0.78-1.0 (diff methods)
- Cerv Jones Pts K=0.45 (sx)
- T1 K=0.60-0.75
- Trunk/LE K=0.44
- Lumbar Bone K=0.48-0.98
- Lumbar Soft Ts K=0.40-0.79
- Lumbar TrP K=0.44
- Agree pain L4-L5>L1-3
- But many poor agreement

Inter-Examiners must agree first … or poor kappas



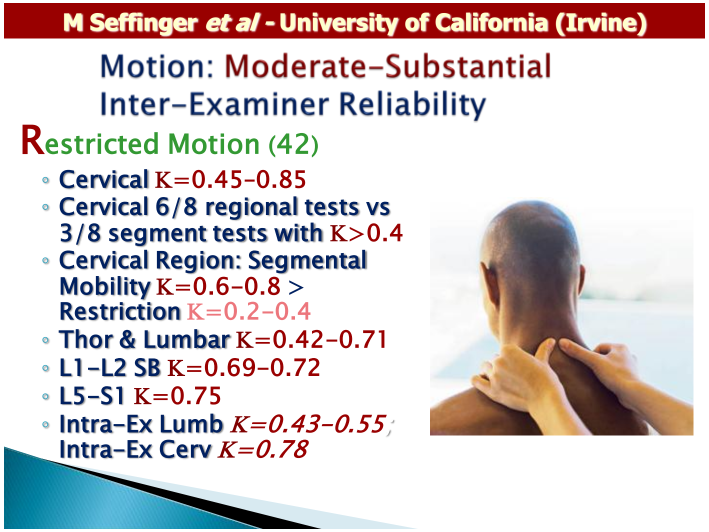
M Seffinger *et al* - University of California (Irvine)

Asymmetry: Substantial Inter-Examiner Reliability

Asymmetry: Landmarks (6)

- *Intra-Exam Lumbar K=0.61-.90*
- Inter-Exam Lumbar K=0.92
- Agreement
  - L4>L1
  - Sit>Prone
  - Some studies no agree!

Inter-Examiners must agree first … or poor kappas

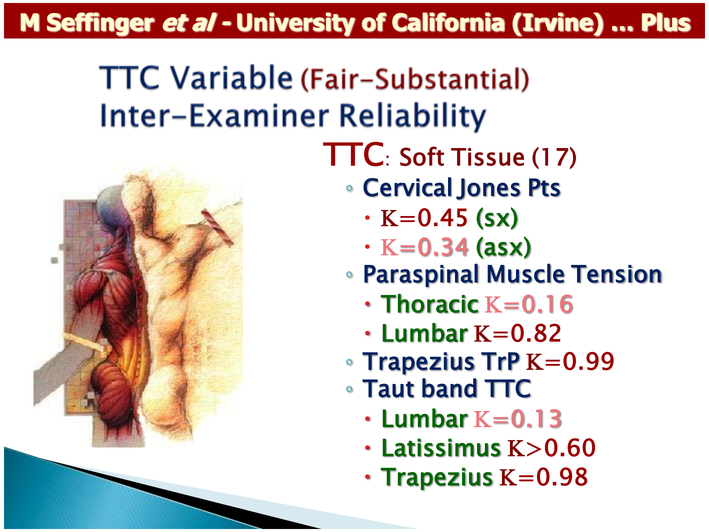


M Seffinger *et al* - University of California (Irvine)

Motion: Moderate-Substantial Inter-Examiner Reliability

Restricted Motion (42)

- Cervical K=0.45-0.85
- Cervical 6/8 regional tests vs 3/8 segment tests with K>0.4
- Cervical Region: Segmental Mobility K=0.6-0.8 > Restriction K=0.2-0.4
- Thor & Lumbar K=0.42-0.71
- L1-L2 SB K=0.69-0.72
- L5-S1 K=0.75
- Intra-Ex Lumb *K=0.43-0.55*; Intra-Ex Cerv *K=0.78*



M Seffinger *et al* - University of California (Irvine) … Plus

TTC Variable (Fair-Substantial) Inter-Examiner Reliability

TTC: Soft Tissue (17)

- Cervical Jones Pts
  - K=0.45 (sx)
  - K=0.34 (asx)
- Paraspinal Muscle Tension
  - Thoracic K=0.16
  - Lumbar K=0.82
- Trapezius TrP K=0.99
- Taut band TTC
  - Lumbar K=0.13
  - Latissimus K>0.60
  - Trapezius K=0.98

Expanding the Evidence Base

Example: Inter-Examiner Studies in *JAOA* 104(8):337-52; Aug 2004

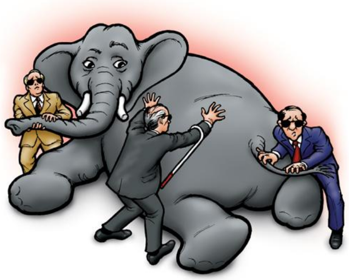
Rivera-Martinez & Capobianco (2 abstracts) static & motion palpation
- L1-5 (K=0.50-0.52)
- T3-7 (K=0.48-0.53)

Driscoll & Friedman *et al* on agreement
- Overall 74% (best lower extremities)
- TART agree 87-95%

Degenhardt *et al:* Inter-Exam asymmetry with camera objective assessment
- Person-person K=0.43-0.74
- Person-Camera several K=0.55-0.67
- Camera-Camera K=0.78

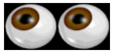
Why a Training Phase? Who Needs One? How to Go About It.

- Students
- Experts
- How Many Needed?
  - Inter-Examiner Reliability of the Test
  - Testing the Teaching of the Test

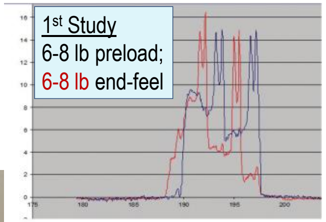


What to Negociate During Training

- Consensus
  - Not too many tests
  - How to do the test step by step (minor details)
    - Position of subject and of examiner
    - Hand positions/ angles / number of repetitions
    - Instructions to subject (if any)
  - The hypothesis: What does the test test? How does it probably work? What is the probable meaning of the test?
  - The judgment: How to report test result (or ambiguity)
- Consider
  - Examiner: Handedness, Dominant Eye
  - Subject: Gender, Body Type, Age



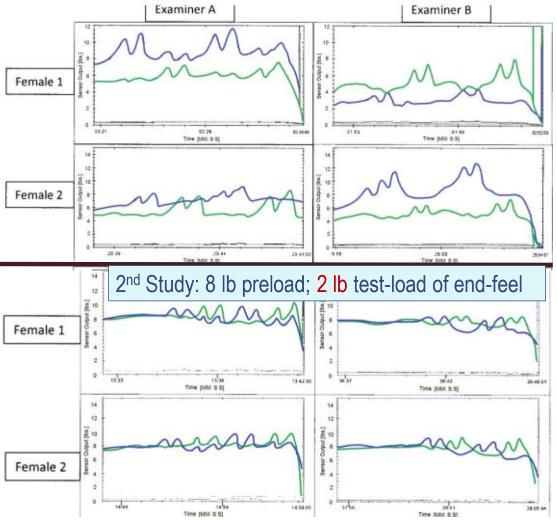
IsoTOUCH® Use in ASIS Compression Test Studies



1st Study
6-8 lb preload;
6-8 lb end-feel

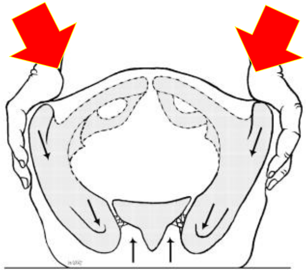
- Tissue Loading Pressures
- Test Pressures
- Negotiations Beyond Pressure

- Asymmetry of pressures right vs left hands intra-examiner & between subjects inter-examiner
- Symmetry of preload and end-feel pressures after **training** with IsoTOUCH® monitors

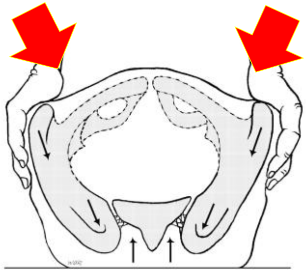


2nd Study: 8 lb preload; 2 lb test-load of end-feel

Angle of the Arms Was Important Too! Direction of Compression to Match SI

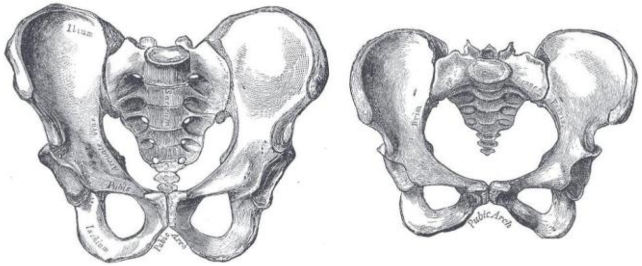


Dominant Eye Center; Side to Stand Height of Table; Foot Position (etc)



Discussion & negotiation ???

Why Might Need to Vary the Test for Male vs Female Pelvises?



Why an Agreement Phase? How Much Agreement is Adequate?

- After training is complete
- Bring in 20 individuals and conduct silent agreement process
- Each examiner makes an evaluation of the 20 consecutive subjects
- If agreement is 80% or better … Conduct your study
- If <80% agreement … Back to the training period for more negotiations towards consensus!

Relation of Overall Agreement to the Prevalance & Obtaining Kappa=0.60

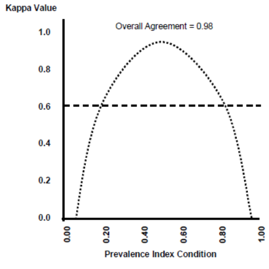


Figure A

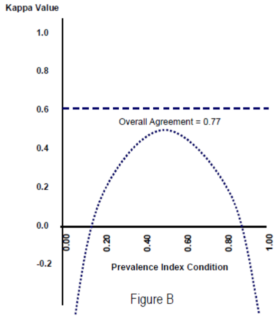
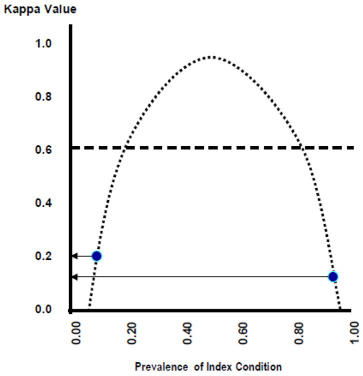


Figure B

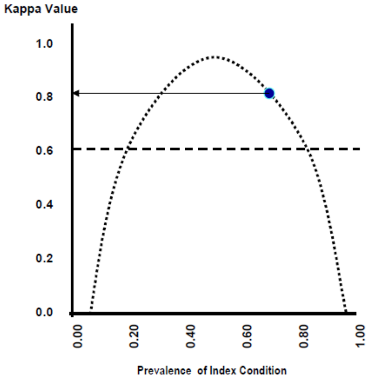
Kappa's Relationship to Prevalence

- Each side of the prevalence bell curve increases the chance that the kappa test will come out poorly



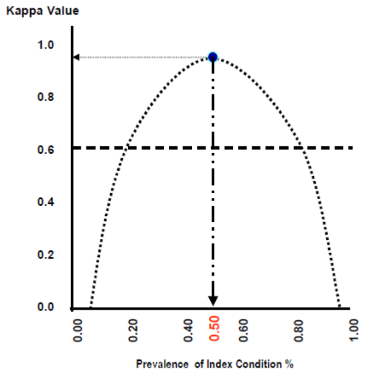
Chances of getting a Kappa=0.60 as Related to Prevalence

- Within certain prevalences of the condition, chances of getting an acceptable kappa increases

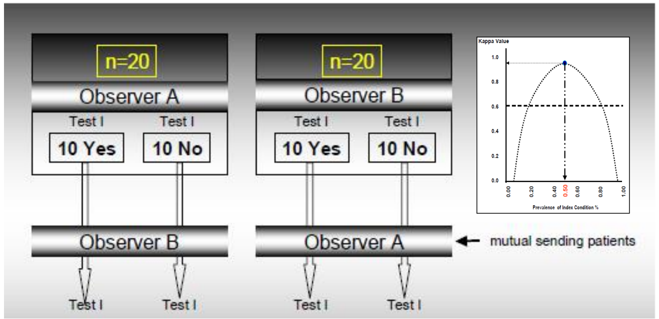


Method to Overcome Prevalence Issue

- Optimum chance of obtaining optimum kappa is when the prevalence is 50%.
- How can you know this in advance?
- How can you recruit study population with as close to 50% prevalence cohort as



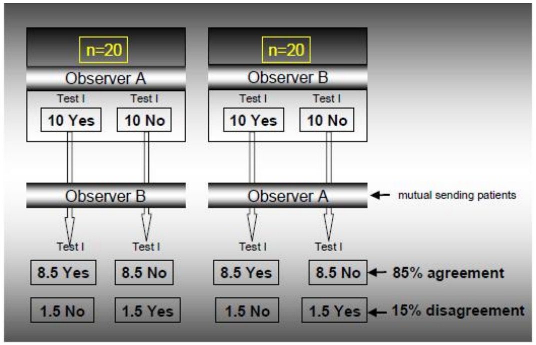
FIMM Protocol to Overcome "Prevalence Pitfall"


n=20
Observer A
Test I
10 Yes
Test I
10 No
Observer B
Test I
Test I

n=20
Observer B
Test I
10 Yes
Test I
10 No
Observer A
Test I
Test I

mutual sending patients

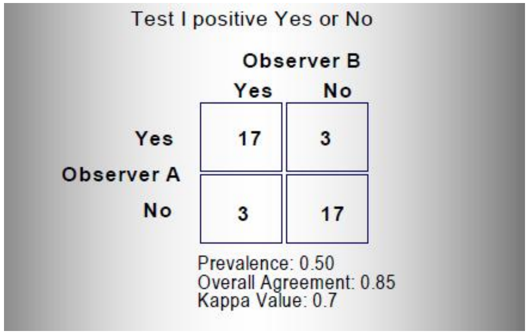
Theoretical Outcome if Adopt FIMM Ground Rules for Testing


n=20
Observer A
Test I
10 Yes
Test I
10 No
Observer B
Test I
8.5 Yes
1.5 No
Test I
8.5 No
1.5 Yes

n=20
Observer B
Test I
10 Yes
Test I
10 No
Observer A
Test I
8.5 Yes
1.5 No
Test I
8.5 No
1.5 Yes

mutual sending patients
85% agreement
15% disagreement

Theoretical Application


Test I positive Yes or No

		Observer B	
		Yes	No
Observer A	Yes	17	3
	No	3	17

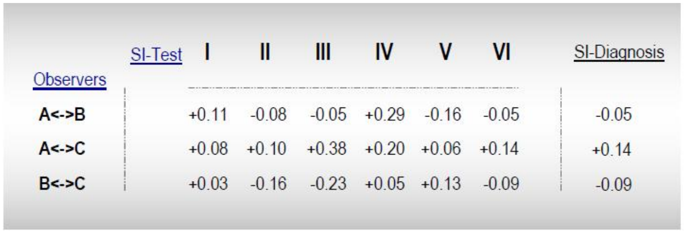
Prevalence: 0.50
Overall Agreement: 0.85
Kappa Value: 0.7

Variation

- In Christian Fossum's prepared lecture, it was suggested that a number of tests be combined to reach a diagnosis
- Few diagnoses are made with a single observation
- How do you test if the tests are independent or not?

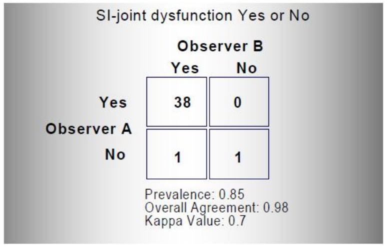
3 Observers with 6 SI Tests Judging SD Based Upon Joint Restriction

- Example:



Observers	SI-Test I	II	III	IV	V	VI	SI-Diagnosis
A<->B	+0.11	-0.08	-0.05	+0.29	-0.16	-0.05	-0.05
A<->C	+0.08	+0.10	+0.38	+0.20	+0.06	+0.14	+0.14
B<->C	+0.03	-0.16	-0.23	+0.05	+0.13	-0.09	-0.09

Same A & B and Tests I, II, III Hypothesis Changed: Test for Muscle Restriction Indicates SD



SI-joint dysfunction Yes or No

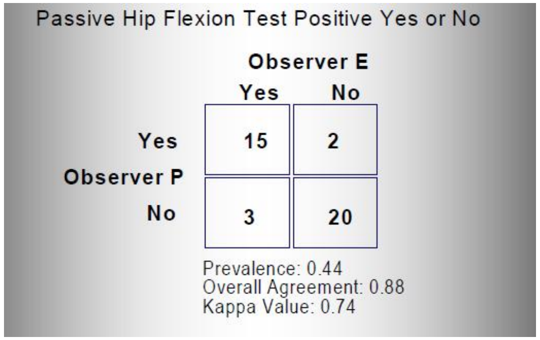
		Observer B	
		Yes	No
Observer A	Yes	38	0
	No	1	1

Prevalence: 0.85
Overall Agreement: 0.98
Kappa Value: 0.7

- Changing the "**hypothesis" of the meaning of the test(s)** changed "absent-to-slight" Kappas to a "substantial" Kappa of 0.70

Actual Outcome: Passive Hip Flexion Test (Patijn, *J Orthop Med*, 2004)

Using FIMM protocol , authors trained, obtained 88% in the agreement phase, and enrolled subject cohort close to 50% prevalence



Passive Hip Flexion Test Positive Yes or No

		Observer E	
		Yes	No
Observer P	Yes	15	2
	No	3	20

Prevalence: 0.44
Overall Agreement: 0.88
Kappa Value: 0.74

Workshop: Training Phase

- Groups of 4-5 (One scribe to write down)
- Pick a diagnostic test (extremity or something seated) – group decision
- Go through Consensus / Training Process
  - Step-by-step how to perform and why
  - Not working? / Not the same? … Negotiate / Compromise
  - Every detail … Side to stand on, how place hands, how many trials, etc
- Group Discussion & Questions

Why Do We Need to Document Palpation & OMT?

- Basic to documenting “somatic dysfunction”
- Documents specifics of how diagnosis made
- Documents what & where we treat with OMT
- Documents if successful when treated (or not)
- Records exactly how SD was treated so that others can replicate same OMT (research articles)
- Expands ability to teach these skills to others

New Data for Inter-Examiner Reliability Evidence-Base

FIMM

- Do/publish more studies
  - Train more to do correctly!
  - Professional leadership
  - By example
- Summary Steps
  - Select test
  - Train, then describe thoroughly
  - Strive for 80%+ “Agreement”
  - Retrain until achieve … renegotiate ambiguity

www.FIMM-Online.com … Scientific Committee

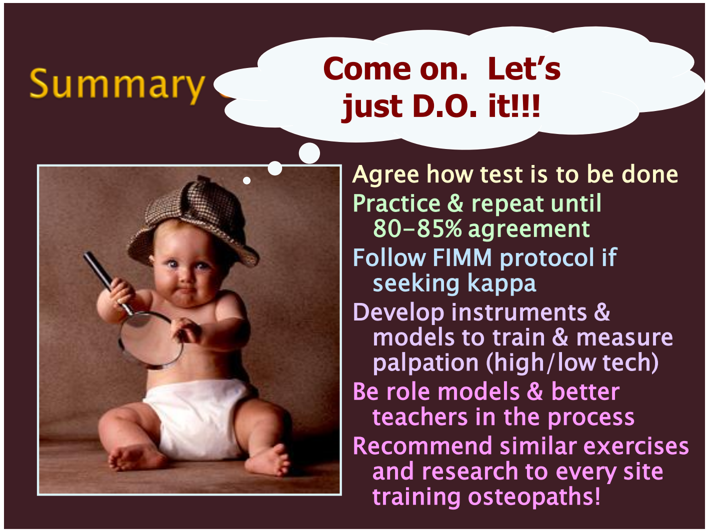
EBM: “Please … Don’t throw the baby out with the bathwater!”

“Simple agreements” = overestimate

“Kappa” measures = underestimate

First studies often poor; learn from mistakes

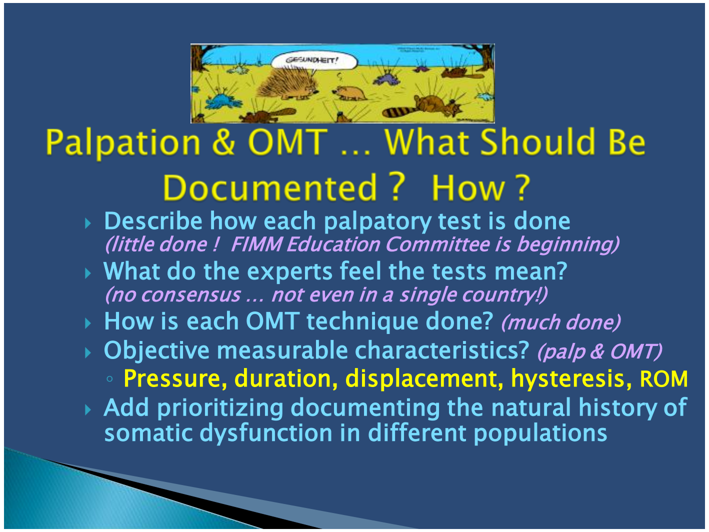
In kappa studies:
- Agreement 1st
- Proper question
- Proper population
- Don’t quit if 1 “poor” outcome



Summary

Come on. Let's just D.O. it!!!

- Agree how test is to be done
- Practice & repeat until 80–85% agreement
- Follow FIMM protocol if seeking kappa
- Develop instruments & models to train & measure palpation (high/low tech)
- Be role models & better teachers in the process
- Recommend similar exercises and research to every site training osteopaths!


GESUNDHEIT!

Palpation & OMT … What Should Be Documented ? How ?

- Describe how each palpatory test is done
  (little done ! FIMM Education Committee is beginning)
- What do the experts feel the tests mean?
  (no consensus … not even in a single country!)
- How is each OMT technique done? *(much done)*
- Objective measurable characteristics? *(palp & OMT)*
  - Pressure, duration, displacement, hysteresis, ROM
- Add prioritizing documenting the natural history of somatic dysfunction in different populations