

Assessing Examiner Agreement

Why Reporting Kappa is NOT Enough

Suggested Reporting Criteria



Background About Kappa

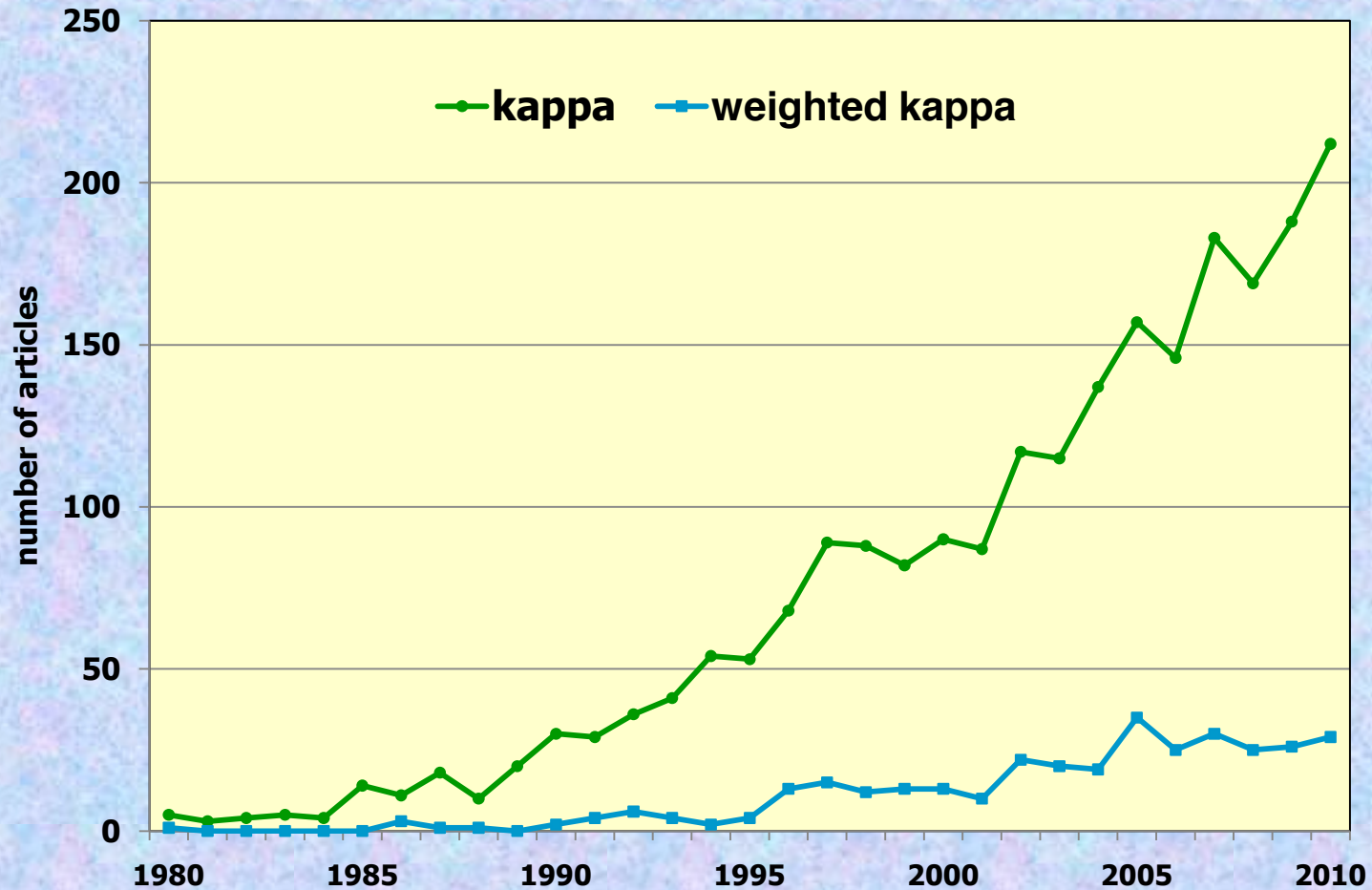
The number of journal articles including some mention of the validity and reliability of new diagnostic methods or investigators employed in clinical studies is increasing, most notably in the past decade.

The Use of Kappa

In particular the use of kappa statistics to assess examiner agreement for categorical outcomes has grown almost exponentially.

A Medline search using "Kappa AND statistic" generated the following.

Frequency of Citations of Kappa (2179 citations)



Incomplete Reporting of Examiner Agreement Using Kappa

The primary purpose of this talk is to demonstrate why reporting only kappa values **does not provide the minimum information needed** to assess examiner proficiency in scoring categorical responses or outcomes.

Incomplete Reporting of Examiner Agreement Using Kappa

The secondary purpose is to suggest different criteria that should be included in reports of examiner proficiency in scoring categorical responses or outcomes.

Incomplete Reporting of Examiner Agreement Using Kappa

Issues regarding the study design (sample size, types of subjects, etc) of the calibration and evaluation of examiner agreement **is not addressed** in this presentation due to time constraints.

Incomplete Reporting of Examiner Agreement Using Kappa

Alternative models to assess examiner agreement, such as log-linear models, latent class models, Bayesian methods and some newer graphical approaches **are not addressed here** either.

Examples of Reporting Kappa

the reliabilities of **three examiners** (weighted kappa) classifying the type of caries treatment needed for occlusal tooth surfaces [none, non-invasive only, or invasive] based on visual versus visual + DIAGNOdent + QLF ranged from **0.50 to 0.61** (Pereira, 2009)

Examples of Reporting Kappa

A good example of reporting kappa involves 10 examiners scoring fluorosis using the TFI. Observed agreement and marginal distributions are presented as well as pairwise kappas (Tavener, 2007)

Is the Focus on Agreement or Validity?

Even when the experts all agree,
they may well be mistaken.

Bertrand Russell

Today we will focus on agreement!

Kappa Statistics

Chance corrected agreement

Simple (Exact) Kappa Statistic

(Cohen, 1960)

$$K = \frac{p_o - p_E}{1 - p_E}$$

Where p_o and p_e represent proportion of observed and expected agreement (under independence, usual X^2 method)

$$-1 \leq -p_E / (1 - p_E) \leq K \leq 1$$

Simple Kappa

A\B	No	Yes	totals
No	40	9	49
Yes	6	45	51
totals	46	54	100

$$p_O = 0.85, \quad p_E = 0.50, \quad K = 0.70$$

How to interpret $K = 0.70$?

Landis and Koch - Biometrics 1977

(cited 8557 times)

Kappa Statistic*

Strength of Agreement

0.81 - 1.00

excellent

0.61 - 0.80

substantial

0.41 - 0.60

moderate

0.21 - 0.40

fair

0.00 - 0.20

slight

< 0.00

poor

*Landis & Koch assumed equal marginals for the examiners

Altman, DG - 1991 Textbook

Kappa Statistic

Strength of Agreement

0.81 - 1.00

very good

0.61 - 0.80

good

0.41 - 0.60

moderate

0.21 - 0.40

fair

< 0.20

poor

Fleiss et al - 2003 Textbook

Kappa Statistic

Strength of Agreement

0.75 - 1.00

very good

0.41 - 0.75

fair to good

< 0.40

poor

Statisticians Confuse the Issue

Kappa = 0.70

Strength of Agreement

Landis-Koch

substantial

Altman

good

Fleiss et al

fair to good

"More" Confusion Regarding Kappa

1. the Kappa statistic is "sufficient" to demonstrate examiner agreement
2. Kappa values found to be "substantial" are generally considered OK.
3. "examiner training ceased when kappa values attained substantial level"

Controversies Regarding Kappa

1. the Kappa value is strongly influenced by the **prevalence** of the outcome
2. Kappa values can be **counter-intuitive**
3. Kappa values can depend on **number** of categories

Controversy 1 - Feinstein, 1990 (prevalence effect)

A\B	No	Yes	totals
No	40	9	49
Yes	6	45	51
totals	46	54	100

$p_o = 0.85$, $K = 0.70$
substantial

A\B	No	Yes	totals
No	80	10	90
Yes	5	5	10
totals	85	15	100

$p_o = 0.85$, $K = 0.32$
fair

Controversy 2 - Feinstein, 1990 (bias has larger kappa than non-bias)

A\B	No	Yes	totals
No	45	15	60
Yes	25	15	40
totals	70	30	100

$p_o = 0.60$, $K = 0.13$
slight

A\B	No	Yes	totals
No	25	35	60
Yes	5	35	40
totals	30	70	100

$p_o = 0.60$, $K = 0.26$
fair

Sources of Examiner Disagreement

In addition to the prevalence effect there are two sources of examiner disagreement

- marginal heterogeneity \approx group level disagreement (marginals)
- bivariate disagreement \approx individual level disagreement (cells) which affects precision

Consequences of Disagreement

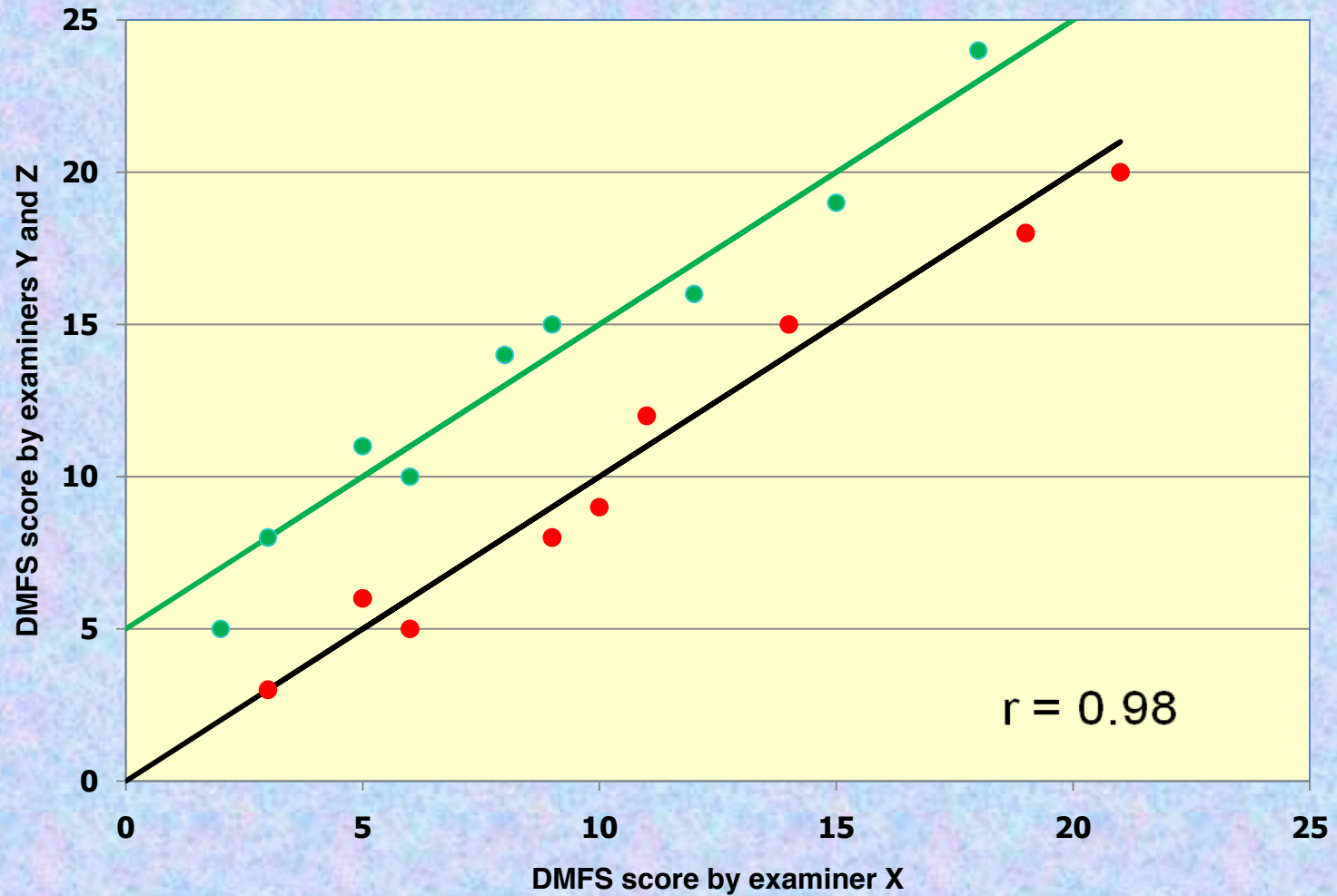
Can produce **biased estimates** of disease prevalence and/or severity.

Can **decrease the power or precision** which **increases the cost** of conducting the study.

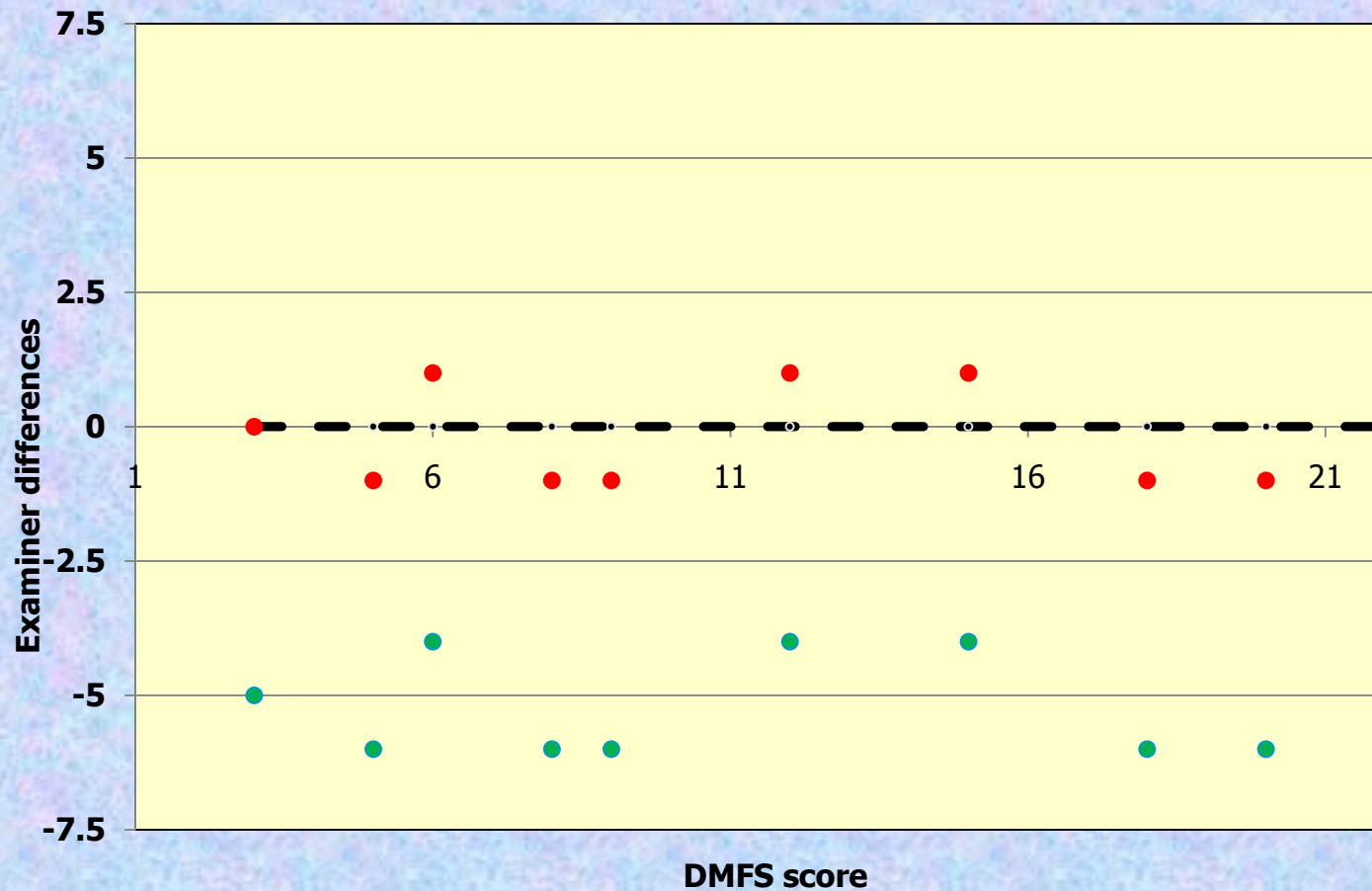
Bias Detection

- Kappa - cannot detect "bias" marginal heterogeneity for categorical variables
- Pearson's correlation coefficient (r) - cannot detect bias for continuous variables.

DMFS Scores for Pairs of Examiners

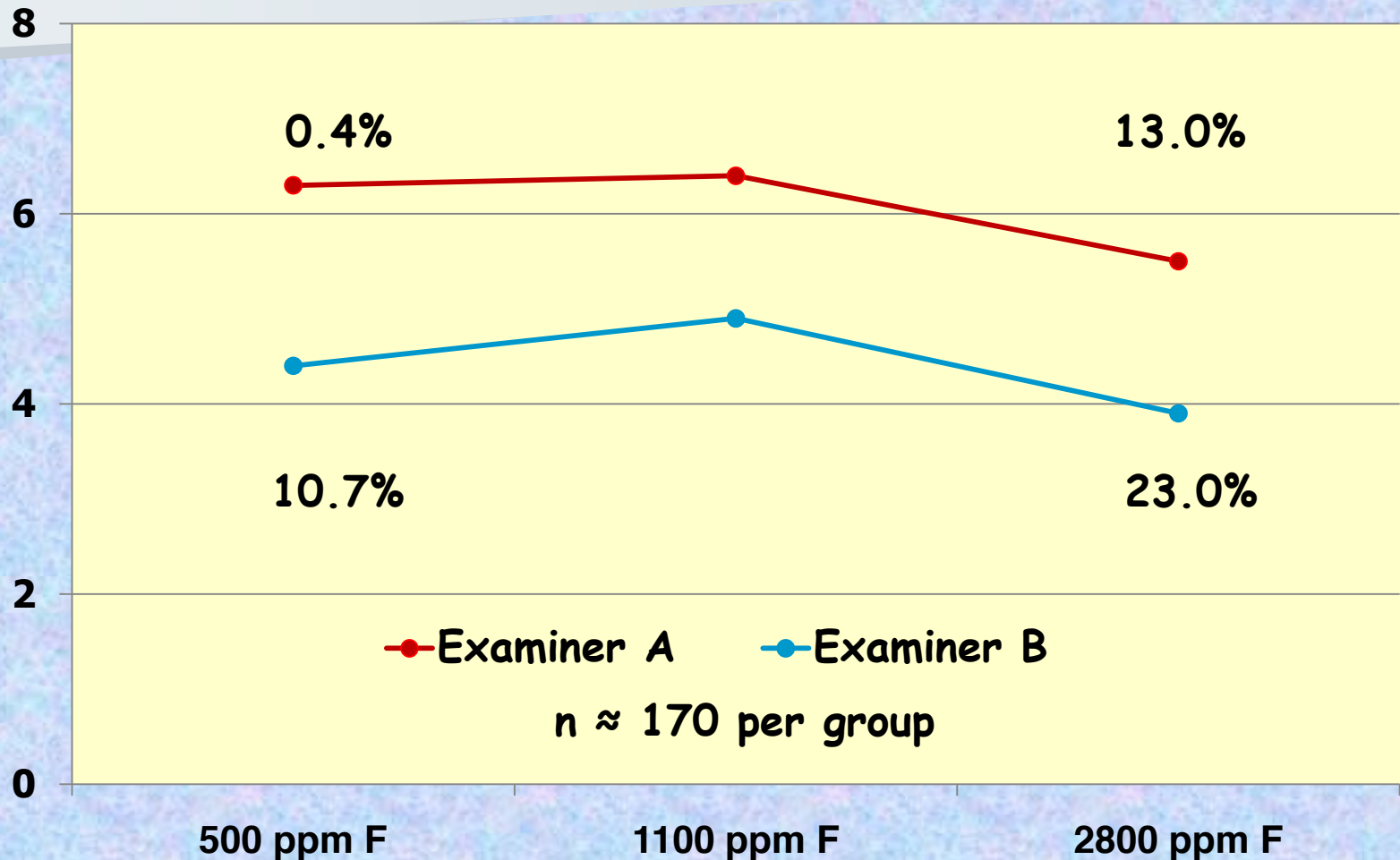


Bland-Altman Differences Plot (cited 18468 times)



Lancet, 1986

Puerto Rico Caries Clinical Trial 24-Month D₂MFS Increments



Stookey, 2004

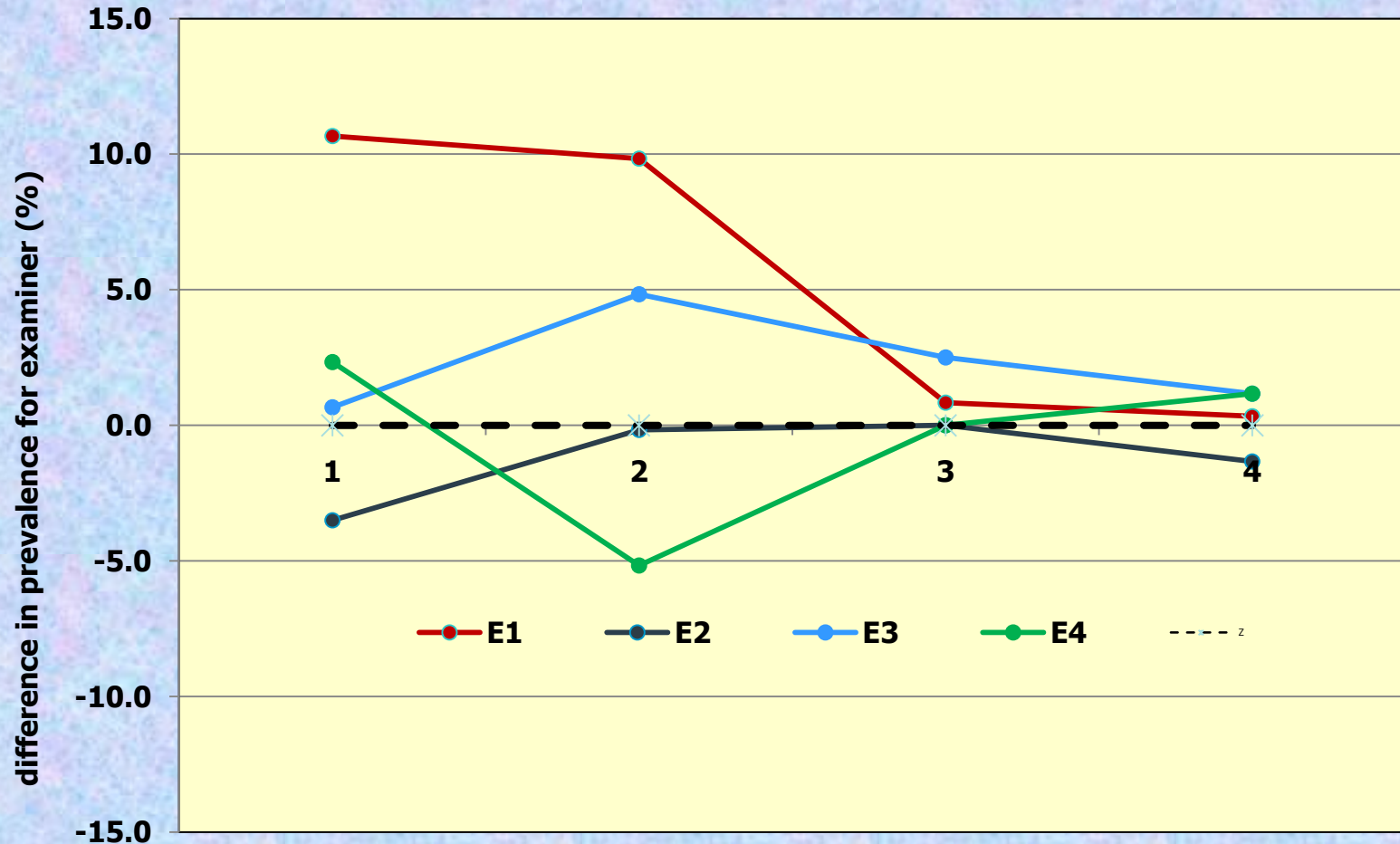
Marginal TFI Distributions for 4 Examiners

(Bias Detection for Ordinal Scale)

TFI	E₁	E₂	E₃	E₄	Avg₁₀
0	35.8	50.0	45.8	44.2	42.4
1	25.8	21.7	20.8	32.5	27.9
2	29.2	20.0	22.5	15.0	21.9
3	7.5	8.3	8.3	5.8	6.9
4	1.7	0.0	2.5	2.5	0.8
n	120	120	120	120	120

Tavener, 2007

Differences Plot in Fluorosis Prevalence



Tavener, 2007

General Kappa (r x r)

$$K = \frac{p_o - p_E}{1 - p_E} = \frac{\sum p_{i.i} - \sum p_{i.}p_{.i}}{1 - \sum p_{i.}p_{.i}}$$

For $r \geq 3$ the range of kappa is

$$-1 < -p_E / (1 - p_E) \leq K \leq 1$$

“Moderate” Agreement (K = 0.60)
No Bias

$X_1 \backslash Y_1$	0	1-2	3-6	Tot X_1
0	158	20	7	185
1-2	18	45	7	70
3-6	5	9	31	45
Tot Y_1	181	74	45	300

$P_O = 0.78$

$P_E = 0.45$

$K = 0.60$

"Moderate" Agreement ($K = 0.60$)

Bias

$X_2 \backslash Y_2$	0	1-2	3-6	Tot X_2
0	145	40	15	200
1-2	6	50	4	60
3-6	4	0	36	40
Tot Y_2	155	90	55	300

$P_O = 0.77$

$P_E = 0.43$

$K = 0.60$

Investigate Marginal Homogeneity (Bias - Group Level)

Bland-Altman Charts - graphic approach

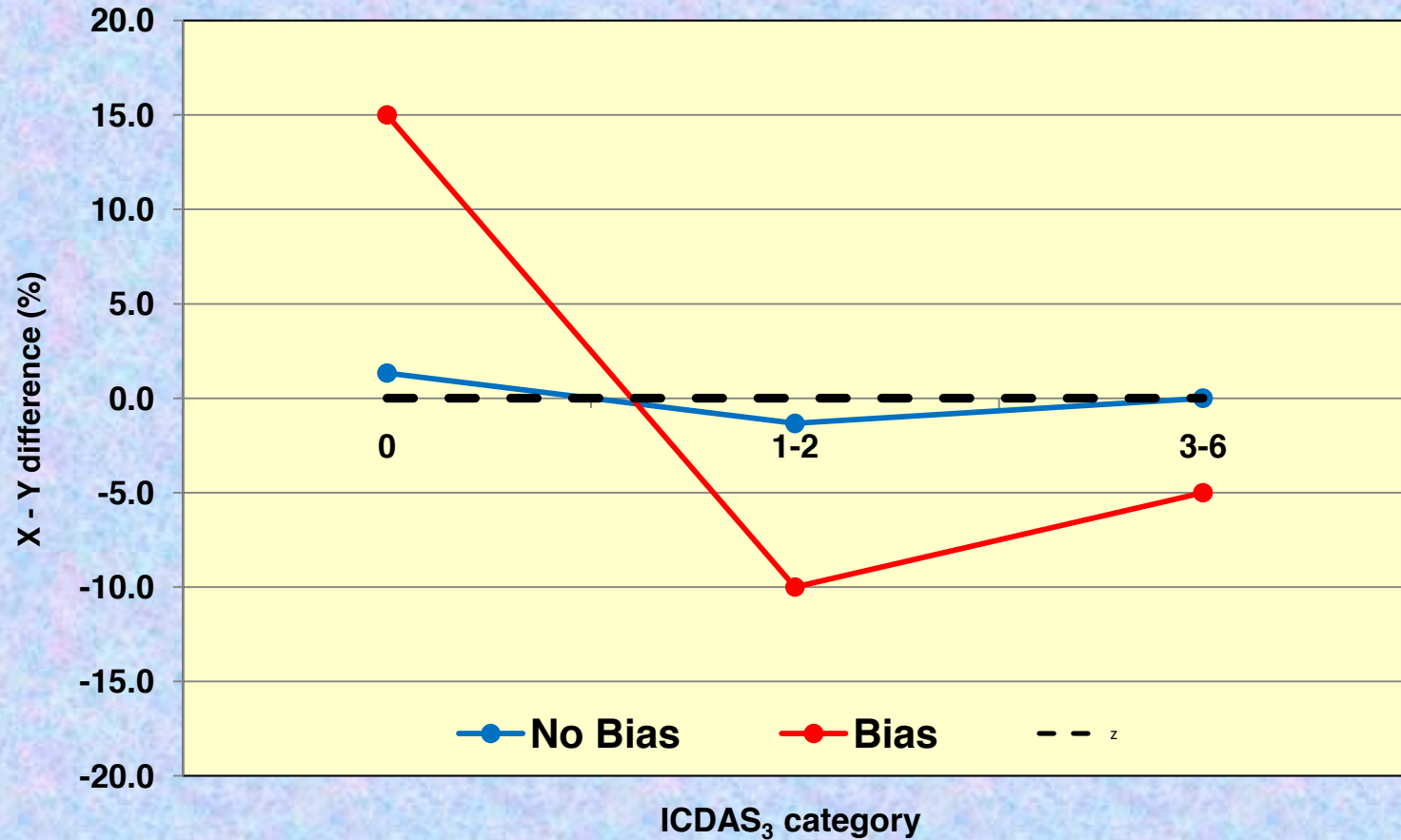
Use maximum kappa - heuristic approach

Perform a X^2 formal statistical test

Marginal Homogeneity?

Category	X₁	Y₁		X₂	Y₂
0	61.7	60.3		66.7	51.7
1-2	23.3	24.7		20.0	30.0
3-6	15.0	15.0		13.3	18.3
N	300	300		300	300

Bland-Altman Marginal Differences Plot



Calculating Maximum Kappa (r x r)

$$K = \frac{p_o - p_E}{1 - p_E} = \frac{\sum p_{i,i} - \sum p_{i.}p_{.i}}{1 - \sum p_{i.}p_{.i}}$$

To maximize kappa we need compute the maximum value for observed agreement, p_o , given fixed marginals. It is derived by maximizing p_o ,

$$\max (p_o) = \sum \min (p_{i.}, p_{.i})$$

Maximum Kappa Approach (no bias)

Category	X₁	Y₁	Min
0	185	181	181
1-2	70	74	70
3-6	45	45	45
sum	300	300	296

$$K = 0.60$$

$$K_M = 0.98$$

Maximum Kappa Approach - (bias)

Category	X₂	Y₂	Min
0	200	155	155
1-2	60	90	60
3-6	40	55	40
sum	300	300	255

$$K = 0.60$$

$$K_M = 0.74$$

Statistical Tests

Marginal Homogeneity Tests

(Sun, 2008)

1. Stuart - Maxwell χ^2 test

$$Z_0 = N \cdot \underline{d}' \bar{V}^{-1} \underline{d}$$

2. Bhapkar χ^2 test

$$Z_1 = \frac{Z_0}{1 - Z_0/N}$$

For $r = 2$ each reduces to McNemar χ^2 test

Marginal Homogeneity Tests

Category	X ₁	Y ₁	d
0	61.7	60.3	1.3
1-2	23.3	24.7	- 1.3
3-6	15.0	15.0	0.0

Bhapkar $X^2 = 0.36$ (df = 2, p = 0.84)

Marginal Homogeneity Tests

Category	X ₂	Y ₂	d
0	66.7	51.7	15.0
1-2	20.0	30.0	- 10.0
3-6	13.3	18.3	- 5.0

Bhapkar $X^2 = 35.0$ (df = 2, p = 0.001)

Tests for Symmetry ($r \times r$)

(Bowker, 1948)

Bowker's X^2 test (in SAS®)

For $r = 2$ it reduces to McNemar X^2 test

X^2 Tests - Symmetry ($K = 0.60$)

X\Y	0	1-2	3-6	Tot X
0	158	20	7	61.7
1-2	18	45	7	23.3
3-6	5	9	31	15.0
Tot Y	60.3	24.7	15.0	300

Bowker $X^2 = 0.69$ (df = 3, p = 0.88)

X^2 Tests - Symmetry ($K = 0.60$)

$X_2 \backslash Y_2$	0	1-2	3-6	Tot X_2
0	145	40	15	66.7
1-2	6	50	4	20.0
3-6	4	0	36	13.3
Tot Y_2	51.7	30.0	18.3	300

Bowker $X^2 = 35.5$ (df = 3, p = 0.001)

Symmetry \neq Marginal Homogeneity

Symmetry \rightarrow Marginal Homogeneity

Marginal Homogeneity \neq Symmetry

Symmetry vs Marginal Homogeneity (K = 0.25)

X\Y	0	1-2	3-6	Tot X
0	50	50	0	100
1-2	40	30	30	100
3-6	10	20	70	100
Tot Y	100	100	100	300

Bowker $X^2 = 13.1$ (df = 3, p = 0.004)

Bhapkar $X^2 = 0.0$ (df = 2, p = 1.00)

Suggestion - Test for Bias First

If marginal homogeneity is rejected, that is, bias is detected, **examine the table to determine the source of the problem (which categories)** and **provide additional training for the examiners.**

Weighted Kappas

Why Weighted Kappa (r x r)

Kappa treats all disagreements in an $r \times r$ table the same. Appropriate for **nominally scaled** variable (race, ethnic group, college major, opinions, values, mental disorders).

Why Weighted Kappa ($r \times r$)

However, for an **ordinally scaled** variable like ICDAS a disagreement between a score of 1 vs 2 (wet vs dry NC) is not as severe as one between 1 vs 5 (NC vs obvious frank lesion). Assign different weights w_{ij} to off-diagonal cells.

Weighted Kappa Statistic ($r \times r$)

$$K_w = \frac{p_{O(w)} - p_{E(w)}}{1 - p_{E(w)}} = \frac{\sum w_{ij} p_{ij} - \sum w_{ij} p_{i.} p_{.j}}{1 - \sum w_{ij} p_{i.} p_{.j}}$$

w_{ij} - selected so that $0 \leq w_{ij} \leq 1$
(think partial credit)

Common Weighted Kappas

1. Linear weights - weights are

$$w_{ij} = 1 - \frac{|i - j|}{(r - 1)}$$

2. Fleiss-Cohen (intraclass or squared error) -

$$w_{ij} = 1 - \frac{(i - j)^2}{(r - 1)^2}$$

Common Weights for (3 x 3) Table

Linear

X\Y	S	NC	C
S	1	1/2	0
NC	1/2	1	1/2
C	0	1/2	1

Intraclass (FC)

X\Y	S	NC	C
S	1	3/4	0
NC	3/4	1	3/4
C	0	3/4	1

Linear Weighted Kappa ($K = 0.60$)

No Bias

$X_1 \backslash Y_1$	0	1-2	3-6	Tot X_1
0	158	20	7	185
1-2	18	45	7	70
3-6	5	9	31	45
Tot Y_1	181	74	45	300

$P_O = 0.78$

$P_E = 0.45$

$K = 0.60$

$K_L = 0.64$

Intraclass Weighted Kappa ($K = 0.60$)

No Bias

$X_1 \backslash Y_1$	0	1-2	3-6	Tot X_1
0	158	20	7	185
1-2	18	45	7	70
3-6	5	9	31	45
Tot Y_1	181	74	45	300

$P_O = 0.78$

$P_E = 0.45$

$K = 0.60$

$K_{FC} = 0.69$

Intraclass Weighted Kappa ($K = 0.25$)

No Bias

X\Y	0	1-2	3-6	Tot X
0	50	50	0	100
1-2	40	30	30	100
3-6	10	20	70	100
Tot Y	100	100	100	300

$$K = 0.25$$

$$K_L = 0.40$$

$$K_{FC} = 0.55$$

Individual Category Weights (3 x 3)

Exact

X\Y	0	1	2
0	1	0	0
1	0	1	0
2	0	0	1

Category 0

X\Y	0	1	2
0	1	0	0
1	0	1	1
2	0	1	1

Category 1

X\Y	0	1	2
0	1	0	1
1	0	1	0
2	1	0	1

Category 2

X\Y	0	1	2
0	1	1	0
1	1	1	0
2	0	0	1

Interpretations of Weighted Kappas

- **Exact Kappa** - can be expressed or viewed as a weighted average of individual category kappas.

Category Kappas and Exact Kappa (Fleiss et al, 2003)

$$K_0 = 0.65, \quad K_1 = 0.51, \quad K_2 = 0.63$$

$$K = \frac{[p_o(0) - p_e(0)] + [p_o(1) - p_e(1)] + [p_o(2) - p_e(2)]}{[1 - p_e(0)] + [1 - p_e(1)] + [1 - p_e(2)]} = 0.60$$

Interpretations of Weighted Kappas

- **Exact Kappa** - can be expressed or viewed as a weighted average of individual category kappas.
- **Linear kappa** - can be expressed as a weighted average of nested category prevalence kappas.

Prevalence Kappas and Linear Kappa

(Van Belle, 2009)

$$K_{1-6} = 0.65 \quad K_{3-6} = 0.63$$

$$K_L = \frac{0.5\{pr_o(1) + pr_o(2)\} - 0.5\{pr_e(1) + pr_e(2)\}}{1 - 0.5\{pr_e(1) + pr_e(2)\}} = 0.64$$

Interpretations of Weighted Kappas

- **Exact Kappa** - can be expressed or viewed as a weighted average of individual category kappas.
- **Linear kappa** - can be expressed as a weighted average of nested category prevalence kappas.
- **Fleiss-Cohen kappa** - is equivalent to the intraclass correlation coefficient for the integer scaled categories (continuous case)

Summary for ICDAS₃ Examples

category	K _c		
	no bias	bias	NB - ASYM
0	0.65	0.56	0.25
1-2	0.51	0.56	0.00
3-6	0.63	0.71	0.55
K	0.60	0.60	0.25
prevalence	K _p		
1-6	0.65	0.56	0.25
3-6	0.63	0.71	0.55
K_L	0.64	0.62	0.40

General Reporting Suggestions

- N and observed agreement (P_o)
- Table/Graph of Marginal Distributions/
Bland-Altman Plot/ K_M or χ^2 Test
- Kappa and Linear Kappa (s.e. or LCL)
- Range of category specific kappas

Specific Reporting Suggestions

Nominal

N and (P_o)

Table/Graph of Marginal Distributions/
(Bland-Altman Plot), K_M or X^2 test

K and s.e./LCL

Range of CSK's

Ordinal

N and (P_o)

K, K_L and s.e./LCL)

Range of PSK's

Sample Template for 4 Examiners

$X_1 \backslash X_2$	A	B	C	D
A	----	N P_o K_M/X^2	N P_o K_M/X^2	N P_o K_M/X^2
B	K (s.e.) K_L (s.e.)	----	N P_o K_M/X^2	N P_o K_M/X^2
C	K (s.e.) K_L (s.e.)	K (s.e.) K_L (s.e.)	----	N P_o K_M/X^2
D	K (s.e.) K_L (s.e.)	K (s.e.) K_L (s.e.)	K (s.e.) K_L (s.e.)	----

Sample Template for 4 Fluorosis Examiners

$X_1 \backslash X_2$	1	2	3	4
1	----	120 62% B	120 54% B	120 69% B
2	0.59 (0.06)	----	120 77% NB	120 70% NB
3	0.52 (0.06)	0.77 (0.07)	----	120 59% B
4	0.64 (0.06)	0.67 (0.07)	0.56 (0.06)	----

Tavener, 2007



Her portrait by Picasso

Near the end of her life,
Gertrude Stein inquired of her
close friend,

“Well, what is the answer?”

Alice B. Toklas remained silent.

Ms. Stein probed, “In that case,
what’s the question?”

As we plod onward thru the fog!

Thank you!

San Giamano, Italy

References

1. Pereira AC. et al (2009). Validity of caries detection on occlusal surfaces and treatment decisions. *Eur J Oral Sci.* 117:51-57.
2. Tavener J, Davies R, Ellwood R (2007) Agreement amongst examiners assessing dental fluorosis from digital photographs using the TF index. *Comm Dent Health.*24: 21-25.
3. Fleiss J, Levin B & Paik M (2003). Statistical Methods for Rates & Proportions, 3rd Ed. Wiley & Sons, New York.
4. Landis J. & Koch G. (1977) Measurement of Observer Agreement for Categorical Data, *Biometrics.* 33: 159-174.
5. Altman, DG. (1991). Practical Statistics for Medical Research, Chapman & Hall, London.

References

6. Feinstein A & Cicchetti D. (1990) High Agreement Low Kappa I. Problems of two paradoxes. *J Clin Epidemiol.* 43: 543-549.
7. Stookey G, Mau M, Isaacs R et al. (2004) The relative effectiveness of 3 fluoride dentifrices in Puerto Rico, *Caries Res* 38: 542-550.
8. Bland J. & Altman D. (1986) Statistical Methods for assessing agreement between 2 Methods of Clinical Measurement. *Lancet* 327: 307-310.
9. Cohen J. (1960) A Coefficient of Agreement for Nominal Scales *Educ & Psychological Measurement* 20: 37-46.
10. Sun X & Yang Z (2008) Generalized McNemar's Test for Homogeneity of Marginal Distributions - paper 382-2008 <http://www2.sas.com/proceedings/forum2008/TOC.html>

References

11. Bowker AH. (1948) A test for symmetry in contingency tables. *J Amer Stat Assoc* 43(244): 572-574.
12. Vanbelle S & Albert A (2009) A note on the linearly weighted kappa coefficient for ordinal scales. *Stat Method* 6: 157-163.