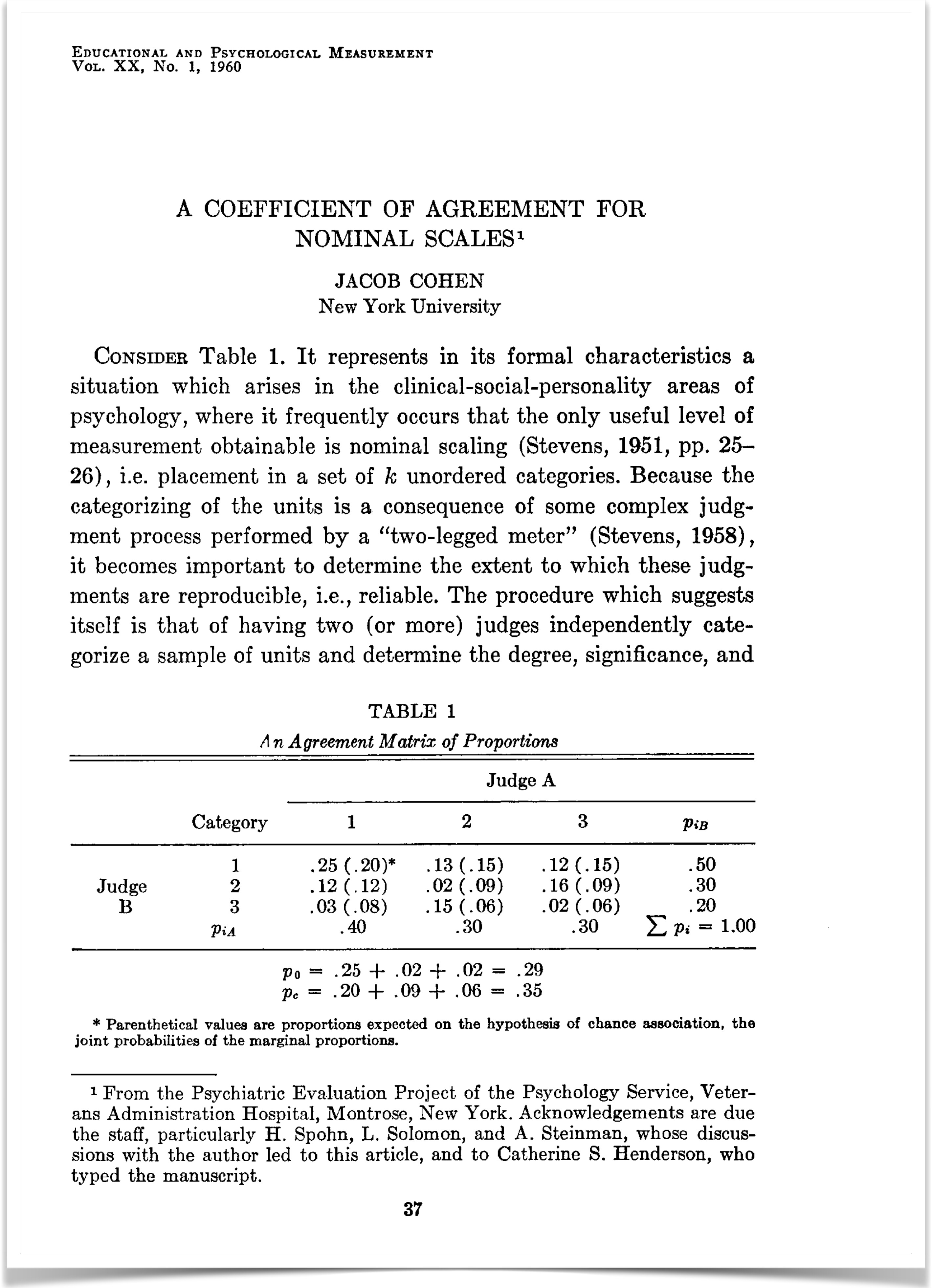
Kappa

Prof. Dr. Ricardo Primi

USF

UNIVERSIDADE SÃO FRANCISCO



EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
VOL. XX, No. 1, 1960

A COEFFICIENT OF AGREEMENT FOR NOMINAL SCALES[1]

JACOB COHEN
New York University

CONSIDER Table 1. It represents in its formal characteristics a situation which arises in the clinical-social-personality areas of psychology, where it frequently occurs that the only useful level of measurement obtainable is nominal scaling (Stevens, 1951, pp. 25–26), i.e. placement in a set of k unordered categories. Because the categorizing of the units is a consequence of some complex judgment process performed by a "two-legged meter" (Stevens, 1958), it becomes important to determine the extent to which these judgments are reproducible, i.e., reliable. The procedure which suggests itself is that of having two (or more) judges independently categorize a sample of units and determine the degree, significance, and

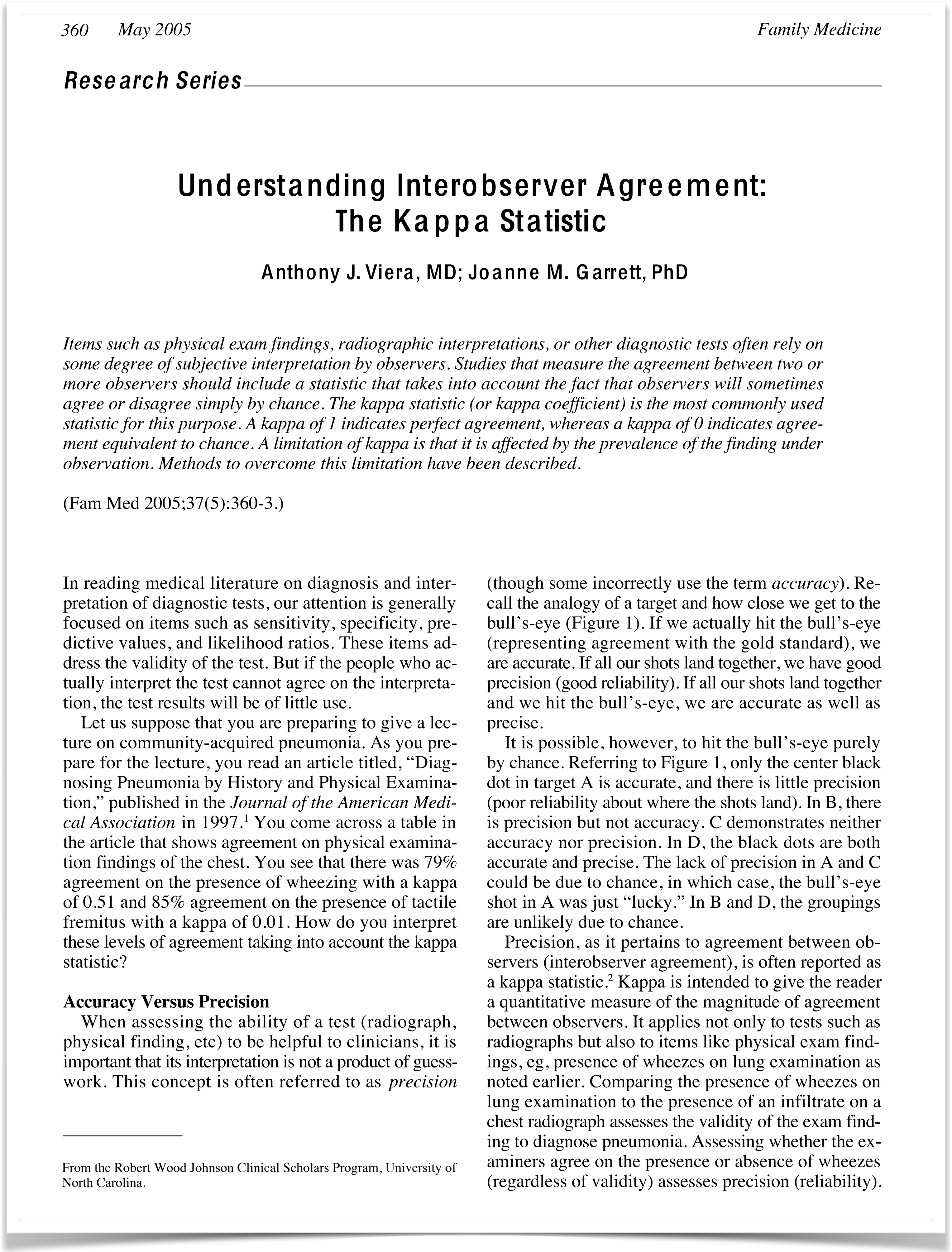
TABLE 1
An Agreement Matrix of Proportions

		Judge A			
	Category	1	2	3	p_{iB}
Judge B	1	.25 (.20)*	.13 (.15)	.12 (.15)	.50
	2	.12 (.12)	.02 (.09)	.16 (.09)	.30
	3	.03 (.08)	.15 (.06)	.02 (.06)	.20
	p_{iA}	.40	.30	.30	$\sum p_i = 1.00$

$$p_0 = .25 + .02 + .02 = .29$$
$$p_c = .20 + .09 + .06 = .35$$

* Parenthetical values are proportions expected on the hypothesis of chance association, the joint probabilities of the marginal proportions.

[1] From the Psychiatric Evaluation Project of the Psychology Service, Veterans Administration Hospital, Montrose, New York. Acknowledgements are due the staff, particularly H. Spohn, L. Solomon, and A. Steinman, whose discussions with the author led to this article, and to Catherine S. Henderson, who typed the manuscript.



Research Series

Understanding Interobserver Agreement: The Kappa Statistic

Anthony J. Viera, MD; Joanne M. Garrett, PhD

Items such as physical exam findings, radiographic interpretations, or other diagnostic tests often rely on some degree of subjective interpretation by observers. Studies that measure the agreement between two or more observers should include a statistic that takes into account the fact that observers will sometimes agree or disagree simply by chance. The kappa statistic (or kappa coefficient) is the most commonly used statistic for this purpose. A kappa of 1 indicates perfect agreement, whereas a kappa of 0 indicates agreement equivalent to chance. A limitation of kappa is that it is affected by the prevalence of the finding under observation. Methods to overcome this limitation have been described.

(Fam Med 2005;37(5):360-3.)

In reading medical literature on diagnosis and interpretation of diagnostic tests, our attention is generally focused on items such as sensitivity, specificity, predictive values, and likelihood ratios. These items address the validity of the test. But if the people who actually interpret the test cannot agree on the interpretation, the test results will be of little use.

Let us suppose that you are preparing to give a lecture on community-acquired pneumonia. As you prepare for the lecture, you read an article titled, "Diagnosing Pneumonia by History and Physical Examination," published in the *Journal of the American Medical Association* in 1997.[1] You come across a table in the article that shows agreement on physical examination findings of the chest. You see that there was 79% agreement on the presence of wheezing with a kappa of 0.51 and 85% agreement on the presence of tactile fremitus with a kappa of 0.01. How do you interpret these levels of agreement taking into account the kappa statistic?

Accuracy Versus Precision

When assessing the ability of a test (radiograph, physical finding, etc) to be helpful to clinicians, it is important that its interpretation is not a product of guesswork. This concept is often referred to as *precision* (though some incorrectly use the term *accuracy*). Recall the analogy of a target and how close we get to the bull's-eye (Figure 1). If we actually hit the bull's-eye (representing agreement with the gold standard), we are accurate. If all our shots land together, we have good precision (good reliability). If all our shots land together and we hit the bull's-eye, we are accurate as well as precise.

It is possible, however, to hit the bull's-eye purely by chance. Referring to Figure 1, only the center black dot in target A is accurate, and there is little precision (poor reliability about where the shots land). In B, there is precision but not accuracy. C demonstrates neither accuracy nor precision. In D, the black dots are both accurate and precise. The lack of precision in A and C could be due to chance, in which case, the bull's-eye shot in A was just "lucky." In B and D, the groupings are unlikely due to chance.

Precision, as it pertains to agreement between observers (interobserver agreement), is often reported as a kappa statistic.[2] Kappa is intended to give the reader a quantitative measure of the magnitude of agreement between observers. It applies not only to tests such as radiographs but also to items like physical exam findings, eg, presence of wheezes on lung examination as noted earlier. Comparing the presence of wheezes on lung examination to the presence of an infiltrate on a chest radiograph assesses the validity of the exam finding to diagnose pneumonia. Assessing whether the examiners agree on the presence or absence of wheezes (regardless of validity) assesses precision (reliability).

From the Robert Wood Johnson Clinical Scholars Program, University of North Carolina.

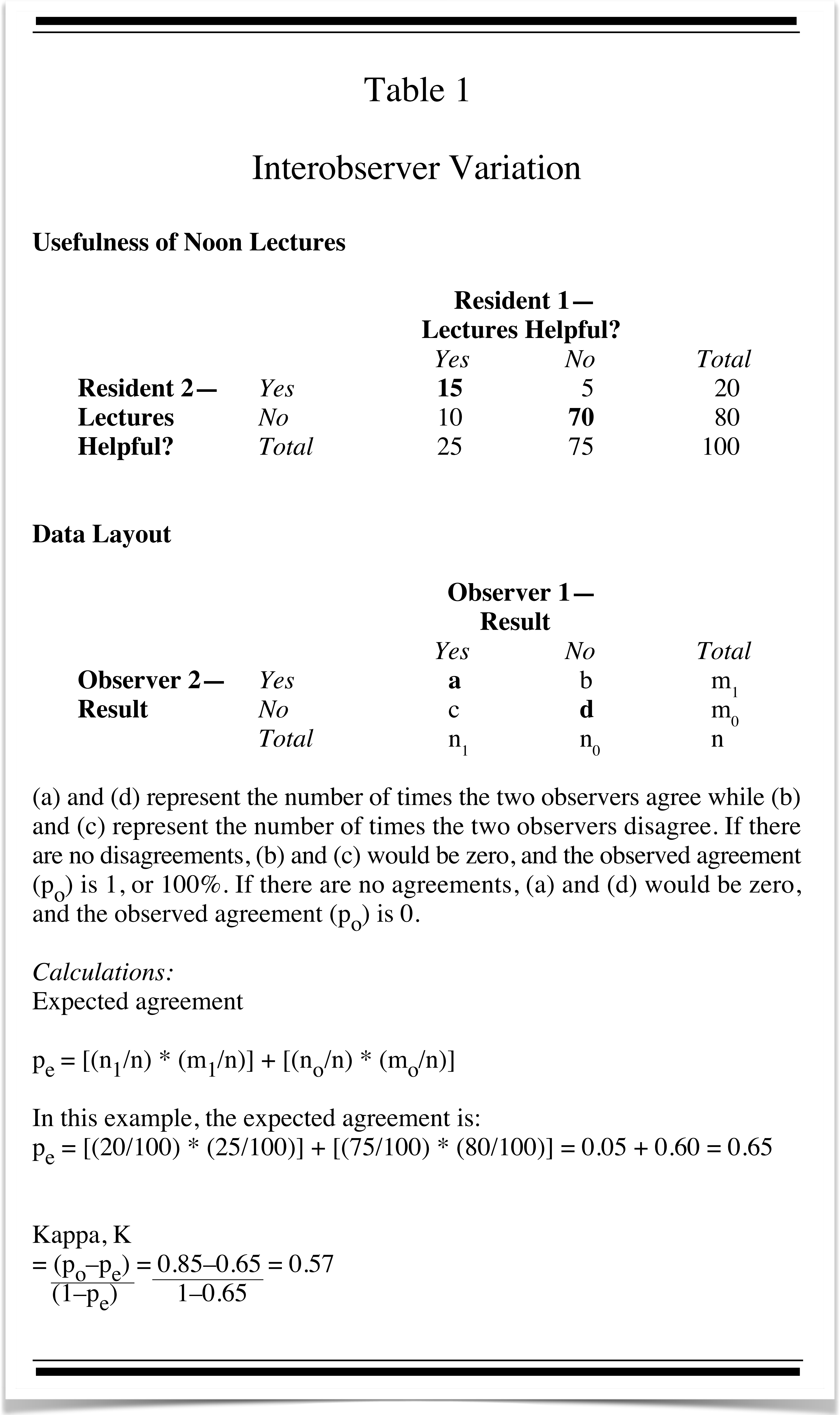


Table 1

Interobserver Variation

Usefulness of Noon Lectures

		Resident 1— Lectures Helpful?		
		Yes	*No*	*Total*
Resident 2—	*Yes*	**15**	5	20
Lectures	*No*	10	**70**	80
Helpful?	*Total*	25	75	100

Data Layout

		Observer 1— Result		
		Yes	*No*	*Total*
Observer 2—	*Yes*	**a**	b	m_1
Result	*No*	c	**d**	m_0
	Total	n_1	n_0	n

(a) and (d) represent the number of times the two observers agree while (b) and (c) represent the number of times the two observers disagree. If there are no disagreements, (b) and (c) would be zero, and the observed agreement (p_o) is 1, or 100%. If there are no agreements, (a) and (d) would be zero, and the observed agreement (p_o) is 0.

Calculations:

Expected agreement

$p_e = [(n_1/n) * (m_1/n)] + [(n_o/n) * (m_o/n)]$

In this example, the expected agreement is:

$p_e = [(20/100) * (25/100)] + [(75/100) * (80/100)] = 0.05 + 0.60 = 0.65$

Kappa, K

$$= \frac{(p_o - p_e)}{(1 - p_e)} = \frac{0.85 - 0.65}{1 - 0.65} = 0.57$$

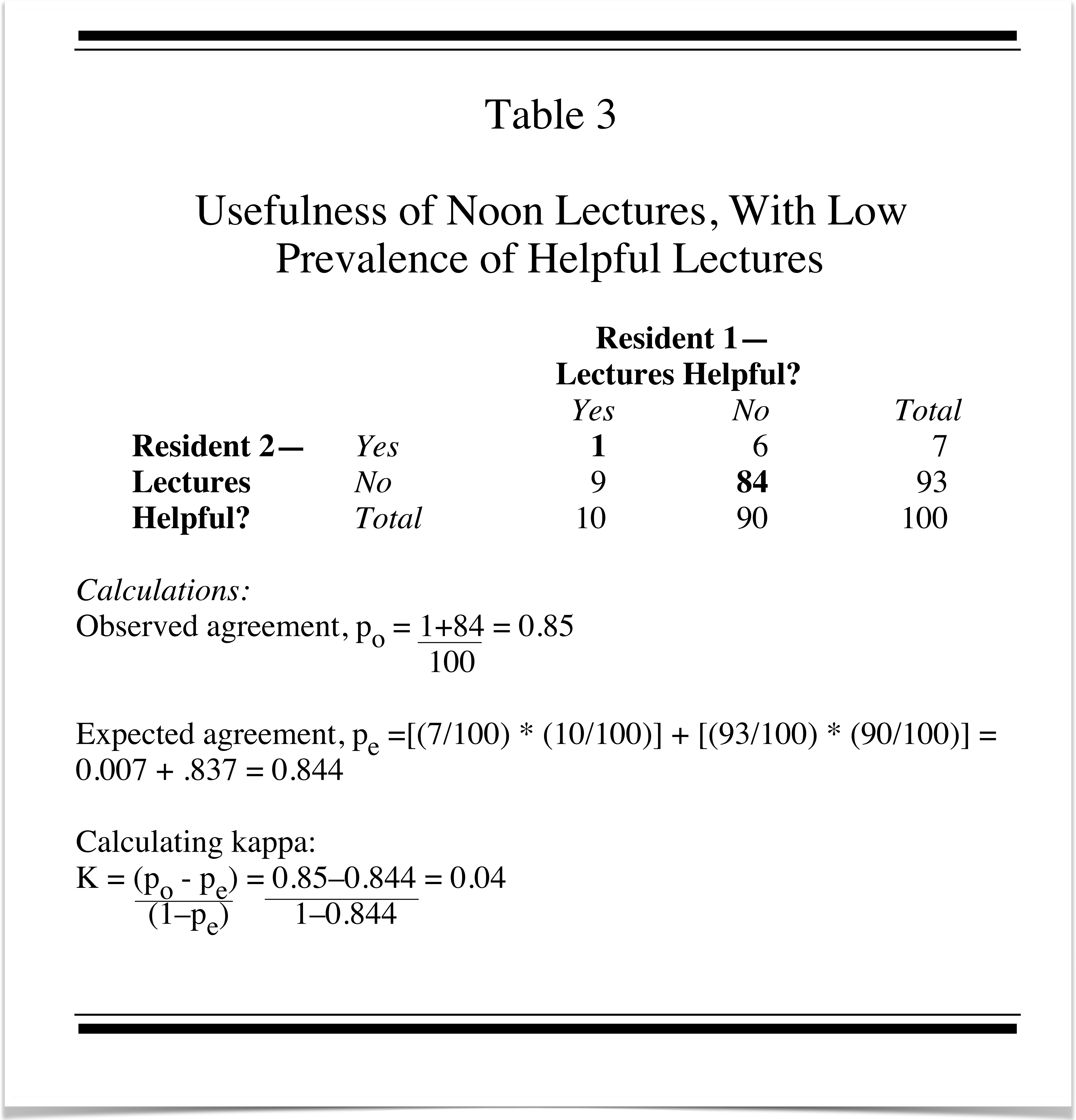


Table 3

Usefulness of Noon Lectures, With Low Prevalence of Helpful Lectures

		Resident 1— Lectures Helpful?		
		Yes	*No*	*Total*
Resident 2—	*Yes*	**1**	6	7
Lectures	*No*	9	**84**	93
Helpful?	*Total*	10	90	100

Calculations:

Observed agreement, $p_o = \frac{1+84}{100} = 0.85$

Expected agreement, $p_e = [(7/100) * (10/100)] + [(93/100) * (90/100)] = 0.007 + .837 = 0.844$

Calculating kappa:

$$K = \frac{(p_o - p_e)}{(1 - p_e)} = \frac{0.85 - 0.844}{1 - 0.844} = 0.04$$

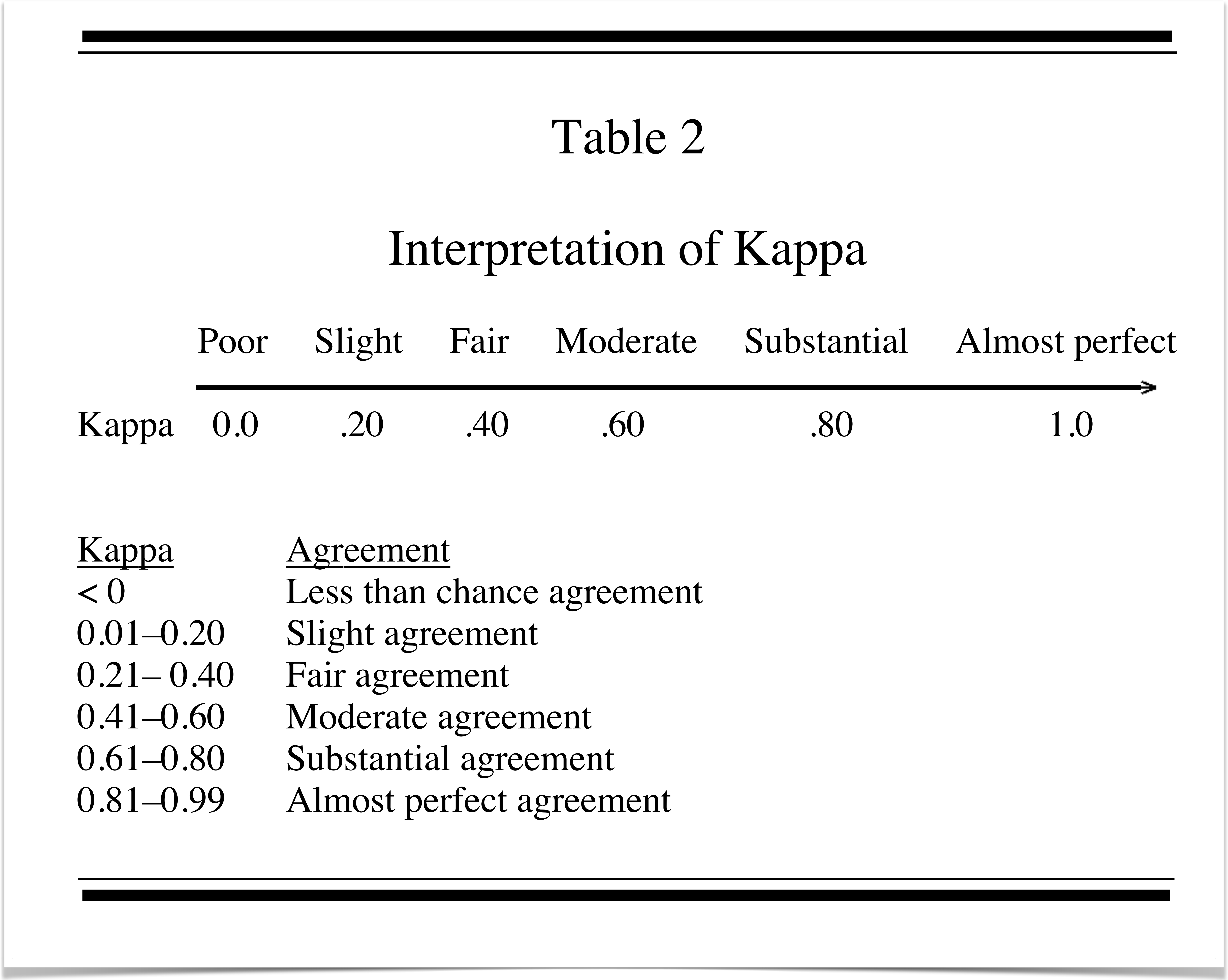


Table 2

Interpretation of Kappa

	Poor	Slight	Fair	Moderate	Substantial	Almost perfect
Kappa	0.0	.20	.40	.60	.80	1.0

Kappa	Agreement
< 0	Less than chance agreement
0.01–0.20	Slight agreement
0.21– 0.40	Fair agreement
0.41–0.60	Moderate agreement
0.61–0.80	Substantial agreement
0.81–0.99	Almost perfect agreement