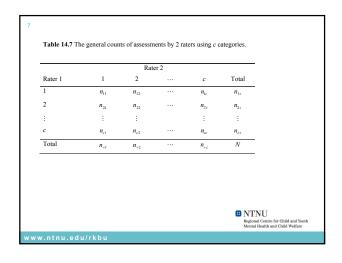
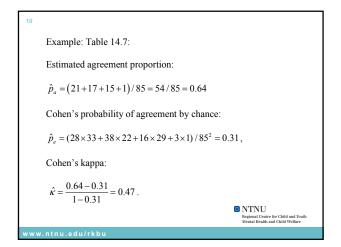


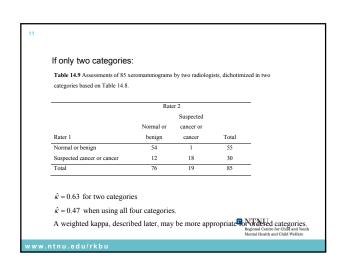
1982).						
		Rat	er 2 Suspected			
Rater 1	Normal	Benign	cancer	Cancer	Total	
Normal	21	12	0	0	33	
Benign	4	17	1	0	22	
Suspected cancer	3	9	15	2	29	
Cancer	0	0	0	1	1	
Total	28	38	16	3	85	

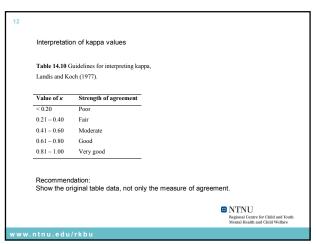


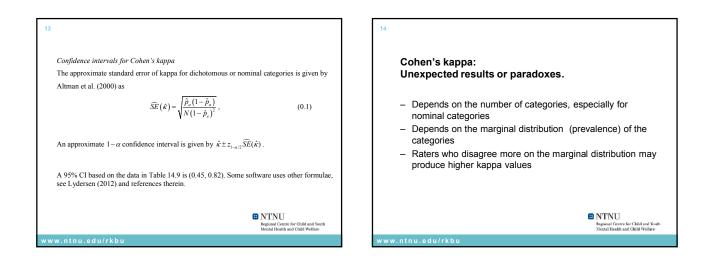
		Rat	er 2			
Rater 1	1	2		с	Total	
1	$p_{11}$	$p_{12}$		$p_{lc}$	$p_{1+}$	
2	$p_{21}$	P <sub>22</sub>		$p_{2c}$	$p_{2+}$	
:	:	:		:	:	
с	$P_{c1}$	$p_{c2}$		$p_{cc}$	$P_{c+}$	
Total	$p_{_{+1}}$	$p_{+2}$		$p_{+c}$	1	

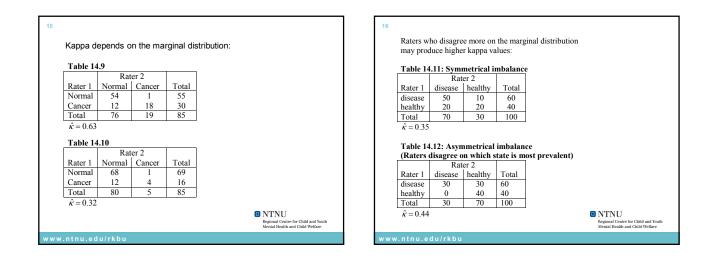
i	Now, consider a situation where two raters each classify subjects in <i>c</i> cate from 1 to <i>c</i> . Let $p_{ij}$ denote the probability that a subject is classified in cate rater 1 and 2, respectively. An intuitive measure of agreement is the probe agree, which is $p_a = p_{11} + p_{22} + + p_{cc}$ .	ergories i and j by
1	But part of this agreement is due to chance. Suppose that rater 1 assigns to probability $p_{i,i} = \sum_{j=1}^{i} p_{ij}$ , and rater 2 assigns to category <i>j</i> with probability independently of rater 1. Then, Cohen's probability of agreement by chan	$p_{+j} = \sum_{i=1}^{c} p_{ij}$
	$p_e = p_{1e}p_{e1} + p_{2e}p_{e2} + + p_{ee}p_{ee} .$ Cohen's kappa is defined as the relative proportion of agreements exceed which is	(0.2) ing that by chance,
www.nt	$\kappa = \frac{p_x - p_y}{1 - p_y}.$	NTN (0.3) Regional Centre for Child and Youth Mental Health and Child Welfare

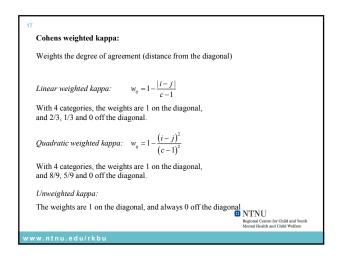


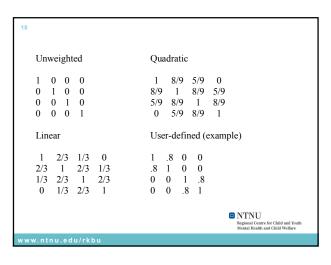


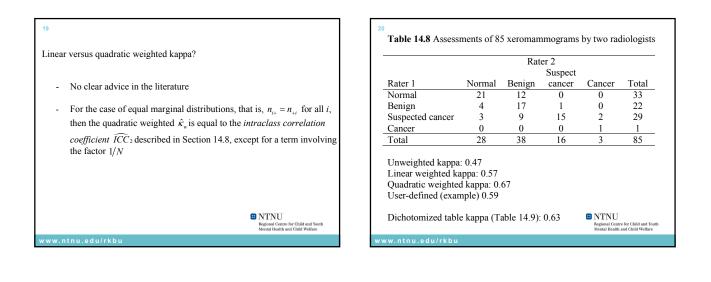


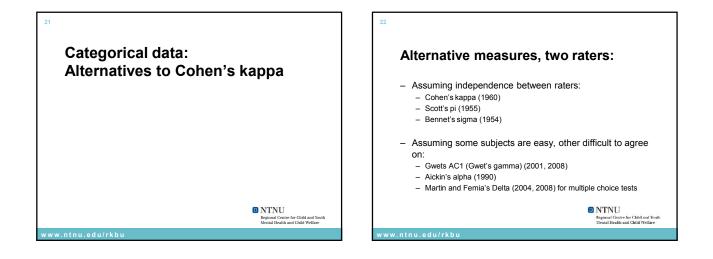


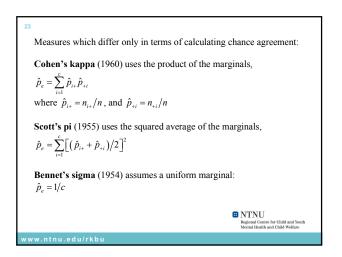


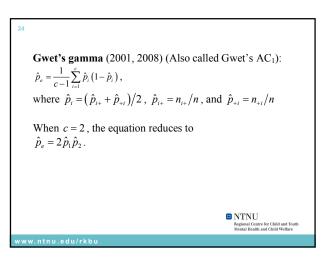




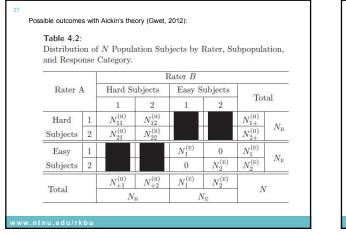




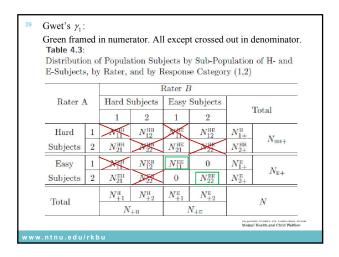


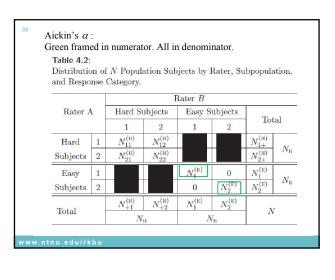


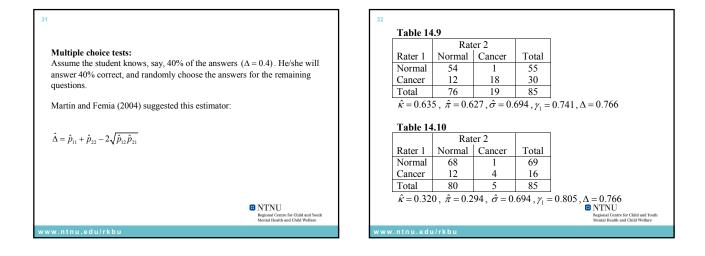
25	26 Possi	ible outcor	nes	with Gwet'	s theory (C	Swet. 201	2):		
Gwet's gamma and Aickin's alpha:	Tal	ble <mark>4.3</mark> :					·		
Easy subjects to classify (E) will be classified (deterministic) in the same category by both raters.		Distribution of Population Subjects by Sub-Population of H- E-Subjects, by Rater, and by Response Category (1,2)			of H- and				
same earegory by boin raters.			1	Rater B					
Hard subjects to classify (H) will be random classified.	ects to classify (H) will be random classified. Rater A			Hard Subjects		Easy Subjects			
Probability $1/c$ for each of the <i>c</i> categories.				1	2	1 2		Total	
Aickin assumes each subject is either hard for both raters (HH), or	I	Hard	1	$N_{11}^{ m HH}$	$N_{12}^{\mathrm{HH}}$	$N_{11}^{\mathrm{HE}}$	$N_{12}^{\text{HE}}$	$N_{1+}^{H}$	N7
easy for both raters (EE).	Su	bjects	2	$N_{21}^{ m HH}$	$N_{22}^{ m HH}$	$N_{21}^{\mathrm{HE}}$	$N_{22}^{\rm HE}$	$N_{2+}^{\mathrm{HH}}$	$N_{ m HH+}$
	1	Easy	1	$N_{11}^{ m EH}$	$N_{12}^{\mathrm{EH}}$	$N_{11}^{\text{EE}}$	0	$N_{1+}^{E}$	N7
Gwet allows also a subject to be hard for Rater 1 and easy for Rater 2 (HE), or vice versa (EH)	Su	bjects	2	$N_{21}^{ m EH}$	$N_{22}^{ m EH}$	0	$N_{22}^{\text{EE}}$	$N_{2+}^{E}$	$N_{\rm E+}$
	-	- + - 1		$N_{\pm 1}^{\mathrm{H}}$	$N_{+2}^{\text{H}}$	$N_{\pm 1}^{\rm E}$	$N_{\pm 2}^{E}$		N
S NTNU	10	Total		$N_{+\mathrm{H}}$		1	$V_{+E}$		11
Regional Centre for Child and Youth Mental Health and Child Welfare	52							Menta	l Health and Child Welfare
vww.ntnu.edu/rkbu	www.ntn	u.edu/	r k b						



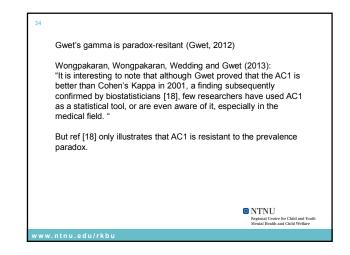
The inter-rater reliability measures (to be estimated) can be expressed as below. These expressions are definitional, since  $N_{\mu}^{EE}$  etc are not observed. Gwet's gamma:  $\gamma_{1} = \frac{\sum_{i=1}^{c} N_{\mu}^{EE}}{N - \left(\sum_{i=1}^{c} N_{\mu}^{HH} + \sum_{i=1}^{c} N_{\mu}^{HE} + \sum_{i=1}^{c} N_{\mu}^{EH}\right)}$ Aickin's alpha:  $\alpha = \frac{\sum_{i=1}^{c} N_{\mu}^{EE}}{N}$ Prove the estimated of the estimat

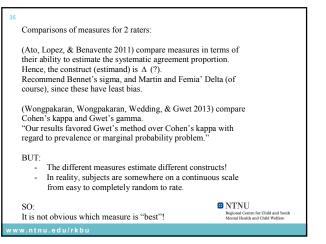


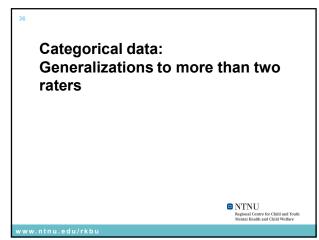




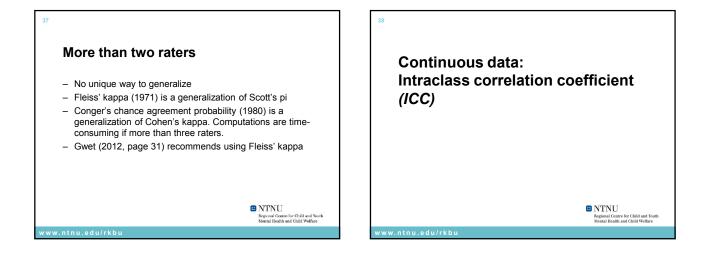
	Rat	er 2					
Rater 1	disease	healthy	Total				
disease	50	10	60				
healthy	20	20	40				
Total	70	30	100				
		nmetrical on which s			lent)		
					lent)		
	disagree o	nmetrical on which s er 2			lent)		
	disagree o	n which s			lent)		
(Raters	disagree o Rat	n which s er 2	tate is m		lent)		
(Raters ) Rater 1	disagree o Rat disease	n which s er 2 healthy	<b>tate is m</b> Total		lent)		
(Raters of Rater 1 disease healthy Total	disagree o Rat disease 30 0 30	<b>n which s</b> er 2 healthy 30 40 70	tate is m Total 60 40 100	ost preva	,		
(Raters of Rater 1 disease healthy Total	disagree o Rat disease 30 0 30	er 2 healthy 30 40	tate is m Total 60 40 100	ost preva	,	00 (or	0.585)

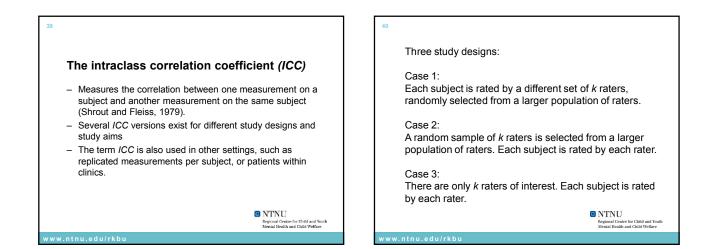






6

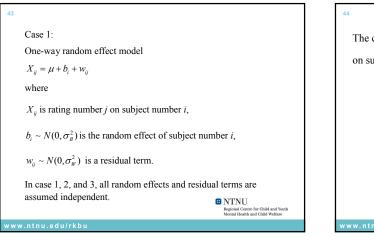


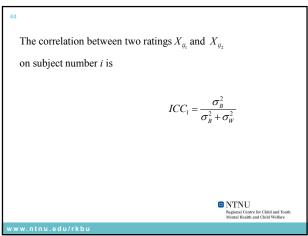


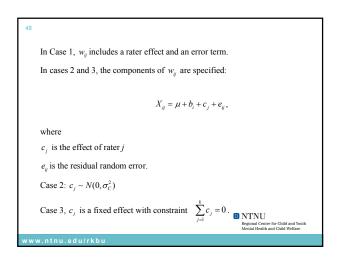
uthors. The IC	C measures w	CC with different model ith k in parentheses are o single measurements.		
	Interaction		Authors	
ANOVA model	between rater and subject?	Shrout and Fleiss (1979)	McGraw and Wong (1996)	Barnhart e al. (2007)
One-way random effects		Case 1 ICC(1,1) or $ICC(1,k)$	Case 1 $ICC(1)$ or $ICC(k)$	ICC <sub>1</sub>
Two-way random effects	Without interaction	As below	Case 2A ICC(A, 1) or $ICC(A, k)$	ICC <sub>2</sub>
	With interaction	Case 2 ICC(2,1) or $ICC(2,k)$	Case 2 As above	ICC <sub>3</sub>
Two-way mixed effects	Without interaction	As below	Case 3A ICC(A, 1) or $ICC(A, k)$	ICC <sub>2</sub>
	With interaction	Case 3 $ICC(3,1)$ or $ICC(3,k)$	Case 3 As above	ICC <sub>3</sub>

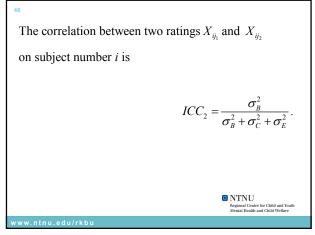
42	
	We shall limit our focus to agreement between single measurements, without interaction, and we use the notation <i>ICC1</i> and <i>ICC2</i> of Barnhart et al. (2007) in Table 14.14.
	Alternatively, agreement can be defined for average of <i>k</i> measurements.
	The intraclass correlation $ICC(3,k)$ in Table 14.14 is equivalent to Cronbach's alpha, a commonly used measure of the internal consistency of items on a psychometric scale.
	INTRU Regional Centre for Child and Youth Mental Hothis and Child Velfare

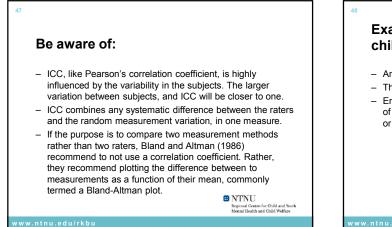
## www.ntnu.edu/rkbu

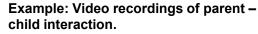












- An RCT of Marte Meo versus treatment as usual
- Three time points: Baseline, 2 months, and 8 months
- Emotional attachment (EA) score based on video recording of parent – child interaction. Rating scored by a psycholigist or psychiatrist.



## Design of Interrater reliability (IRR) study

- 36 distinct individuals, 12 from each of 3 time points.
- Each was rated by 2 raters, from a pool of 4 raters.

References

492.

Ato, M., Lopez, J.J., & Benavente, A. 2011. A simulation study of rater

agreement measures with 2x2 contingency tables. Psicologica, 32, (2) 385-402

Barnhart, H.X., Haber, M.J., & Lin, L.I. 2007. An overview on assessing agreement with continuous measurements. *J.Biopharm.Stat.*, 17, (4) 529-569

Gisev, N., Bell, J.S., & Chen, T.F. 2013. Interrater agreement and interrater reliability: key concepts, approaches, and applications. *Res.Social.Adm Pharm.*, 9, (3) 330-338

Lydersen, S. 2012, "Diagnostic tests, ROC curves, and measures of agreement," In Medical statistics in clinical and epidemiological research, M. B.

Veierød, S. Lydersen, & P. Laake, eds., Oslo: Gyldendal Akademisk, pp. 462-

INTNU

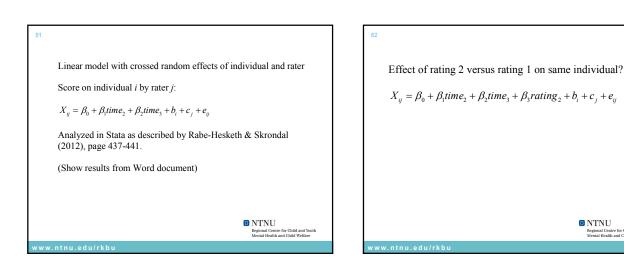
Centre for Child and Youth Tealth and Child Welfare

Gwet, K.L. 2012. Handbook of Inter-rater Reliability, 3 ed. Gaithersburg,

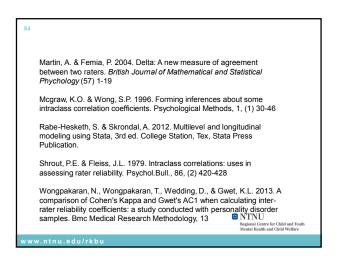
Maryland, Advanced Analytics, LLC.

All 6 combinations of raters rated 2 individuals at each of the 3 time points.

50	<sup>50</sup> Design (continued)							
	<ul> <li>Three first-raters (A, B, C) at each time point.</li> <li>Four second-raters at each time point (A, B, C, D)</li> <li>At each time point 12 pairs of raters.</li> </ul>							
	AD	AB						
	BD	BA						
	CD	BC						
	AD	СВ						
	BD	AC						
	CD	CA	I NTNU					
			Regional Centre for Child and Youth Mental Health and Child Welfare					
www	.ntnu.edu/rk	bu						



NTNU
 Begional Centre for Child and Youth
 Mental Health and Child Welfare



NTNU

ional Centre for Child and Youth ital Health and Child Welfare

## 9