**SUPPLEMENTAL MATERIAL**

***1. Gwet's inter-rater agreement ('kappa') statistic***

This section explains why we chose Gwet's method to measure inter-rater agreement. The following section reviews inter-rater agreement statistics in general terms for readers not familiar with 'kappa' statistics.

A variety of statistical measures of inter-rater agreement, conventionally referred to as 'kappa' (κ) values, have been proposed, including Cohen's kappa for agreement between 2 raters, Fleiss' multi-rater Kappa, and Gwet's multi-rater kappa values (usually called 'agreement coefficients', AC1 for categorical data and AC2 for ordinal data), among others(Gwet, 2010). All such kappa statistics have in common the notion that some portion $PC$ of the observed percentage agreement $PA$ may be due to chance, and thus attempt to estimate the percentage of agreement 'beyond chance' as

$$\kappa = \frac{PA - PC}{100 - PC}$$

i.e. amount of agreement 'left over' after subtracting out that due to chance ($PA - PC$), relative to the maximal possible beyond-chance agreement ($100 - PC$). For readers encountering Kappa statistics for the first time, additional explanation is provided in the next section of this Supplemental material.

The primary distinction between the various available methods for assessing inter-rater agreement lies in the statistical assumptions used in estimating the (unobserved) quantity $PC$, the percentage agreement attributable to chance. Though widely used, Cohen's and Fleiss' $\kappa$ statistics perform badly (exhibit 'paradoxes') when raters exhibit a high or low degree of agreement, and when the true prevalence of classes among the cases being rated is non-uniform(Feinstein and Cicchetti, 1990; Gwet, 2008).

Gwet recently developed improved methods for the multi-rater setting for calculating $PC$ based on a model of chance agreement in which (1) chance agreement occurs when either rater rates a case randomly, and (2) only an portion of the observed ratings (to be estimated) of the observed ratings is subject to randomness, and demonstrated that the resulting $\kappa$ statistic remains 'paradox free' even in high-agreement, high-prevalence settings(Gwet, 2008). The interested reader is referred to the excellent discussion in Chapter 6 of the cited reference(Gwet, 2010).

We used Gwet's multi-rater agreement coefficients AC1 (for categorical data) and AC2 (for ordinal data)(Gwet, 2008; 2010), hereafter referred to simply as kappa ($\kappa$) statistics. We divided the 15 questions for each case into categorical and ordinal assessment types as follows: *ordinal*: sharpness, absolute amplitude, relative amplitude, frequency, number of phases, and evolution; and *categorical*: seizure, main term 1, main term 2, plus modifier, and triphasic morphology. Agreements on categorical assessments were considered to be all-or-none. Partial agreement for ordinal data was scored using a conventional quadratic penalty function, adjusted for the number of possible choices(Gwet, 2010). For example, for evaluating agreement about absolute amplitude, the options are numbered sequentially (very low =1, low =2, medium =3, high=4), and agreement between a pair of raters is given by one minus the squared difference between the numbers assigned by the two raters divided by the squared maximum possible discrepancy. Using this scheme, if two raters scored the amplitude of an EEG feature as 1 (very low) and 2 (medium), respectively, their degree of agreement would be scored as $1 - \frac{(1-3)^2}{(1-4)^2} = 0.55$,

whereas perfect agreement would receive a score of 1, and maximal disagreement would receive a score of zero.

Precision of the group estimates for inter-rater agreement can be quantified by calculating 95% confidence for the estimated Gwet's $\kappa$ values using the Jackknife method(Tukey, 1958; Gwet, 2010).

***2. Inter-rater agreement ('kappa') statistics***

This section provides an informal explanation of Kappa statistics, used to quantify inter-rater agreement. Suppose two EEGers (N & B) evalutate 100 EEGs, and declare each "normal" or (B) "abnormal". Suppose the data are as follows:

| | | N | |
|---|---|---|---|
| | | Normal | Abnormal |
| B | Normal | 50 | 30 |
| | Abnormal | 10 | 10 |

N & B agree on 50+10=60 records, hence the observed % agreement is **a=60/100=60%.** Note that both readers read the majority of records as "normal: B called 80 records (80%) abnormal, and M calls 60 records (60%) abnormal.

Intuitively, some of the agreement in this 60% figure could hypothetically have been due to 'luck', i.e. guessing the same answer without any real expertise. Thus it often makes sense to further assess the % agreement in another way, by correcting for the % agreement attributable to luck.

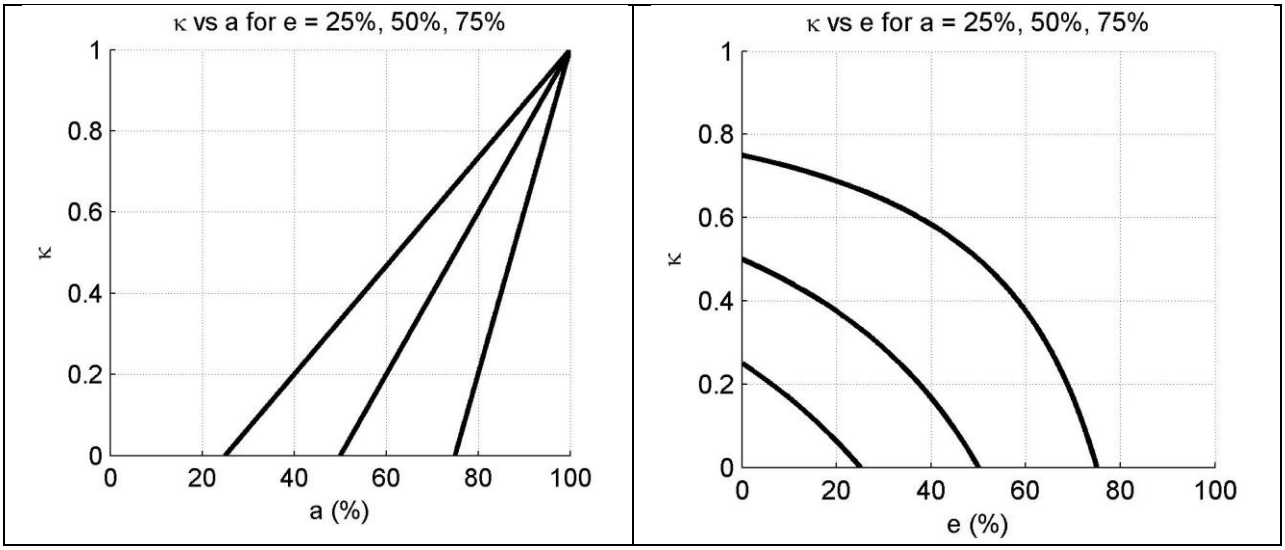
Here is how Kappa statistics adjust for the possibility of chance agreement, or, in other words, evaluate the degree of agreement that exists beyond what is attributable merely to luck. Suppose that, according to some analysis of N & B's answers, we determine that PC=56% of the observed agreement is attributable to chance (the exact value estimated for PC depends on the assumptions made about how chance agreement might arise). Any agreement beyond 56% can be said to have been "real", or "beyond what would be expected by chance". The maximum amount of additional agreement is 100-56 = 44%.

Now consider: What fraction of this additional possible 44% agreement did N & B's EEG readings actually achieve? The answer is simply the additional agreement beyond chance actually observed (i.e. difference between the actual vs hypothetical chance agreement), divided by the total amount of possible additional agreement, i.e.

$$\kappa = \frac{a - e}{1 - e} = \frac{60 - 56}{100 - 56} = \frac{4}{44} = 9.1\%$$

The dependence of the Kappa statistic on the variables $a$ and $e$ can be better understood by studying the plots in **Fig S1**, showing Kappa vs $a$ holding and Kappa vs $e$. Note for example that perfect inter-rater agreement always produces a "perfect Kappa" value ($\kappa = 1$) (leftmost plot). Conversely, the value of $\kappa$ can never exceed the actual observed level of agreement (rightmost plot).

**<u>Supplemental Fig. 1.</u> Plots of Kappa vs the % observed (a) and expected (e) agreements for various values of the expected resp. observed agreement are as follows:**




***<u>3. Characteristics of Participating Clinical Electroencephalographers</u>***

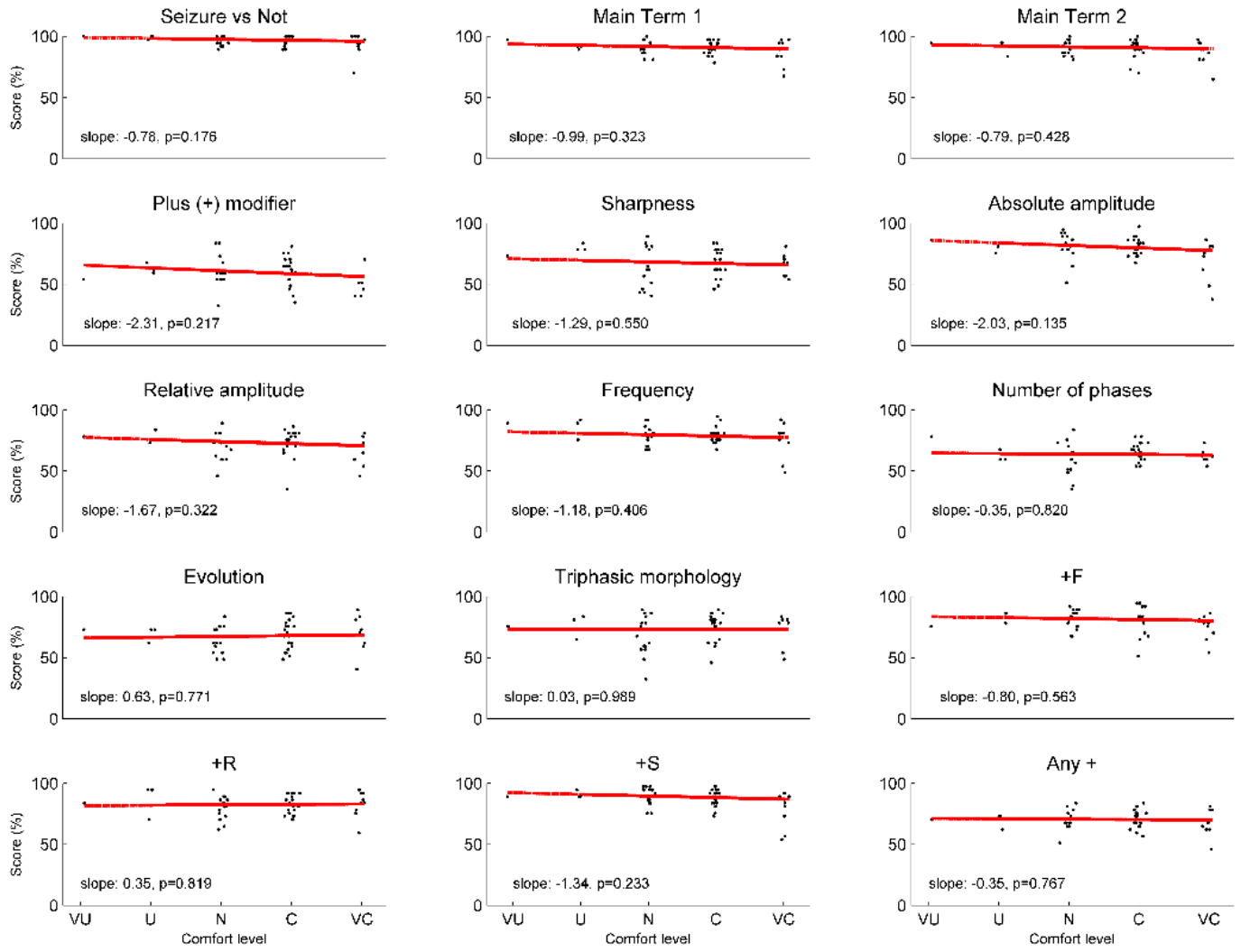
The self-reported levels of involvement in critical care EEG (ccEEG) research, number of years clinical EEG interpretation, and level of comfort with the ACNS ccEEG terminology are summarized in the following table.

**Table S1. Raters' characteristics (N=49)**

| | **N(%)** |
|---|---|
| **Degree of involvement in critical care EEG research** | |
| Principal investigator in the CCEMRC | 15 (31%) |
| Co-investigator in the CCEMRC | 11 (20%) |
| Clinical or research fellow | 20 (41%) |
| Clinical neurophysiologist not involved in research | 3 (6%) |
| **Experience reading EEG (years)** | |
| <2 | 22 (45%) |
| 2-5 | 7 (14%) |
| 5-10 | 10 (20%) |
| 10-15 | 4 (8%) |
| >15 | 6 (12%) |
| **Self-reported comfort with the ACNS Critical Care EEG Terminology** | |
| Very uncomfortable | 2 (4%) |
| Uncomfortable | 3 (6%) |
| Neutral | 16 (33%) |
| Comfortable | 22 (45%) |
| Very comfortable | 7 (14%) |

***4. Lack of statistical relationship between comfort level on test-performance***

This figure shows that there is no strong effect of self-reported comfort level on the score achieved on scores achieved by participants in the IRA study. Scores for all comfort-level groups are statistically indistinguishable. Compare with Fig. 4 in the main text, which shows a similar lack of an effect of years of EEG reading experience.




**Supplemental Fig. 2**: **Effects of comfort level with ACNS Standardized Critical Care EEG Terminology.** Scores (% correct, relative to expert panel consensus answers) were plotted for all raters within five groups of self-reported EEG comfort level with the terminology: (UC, very uncomfortable; U, uncomfortable; N, neutral; C, comfortable; VC, very comfortable). Linear regression lines (red dashed lines, of the form $y = \text{slope} \cdot x + b$), were fitted to each set data, with $x$ values equal to 1,2,3,4 or 5 for the different levels of experience. The slopes and associated $p$-values are shown on each graph**.**