Indexing the Approximate Number System

Matthew Inglis and Camilla Gilmore

Mathematics Education Centre

Loughborough University

Mathematics Education Centre
Loughborough University
Loughborough
Leicestershire, LE11 3TU
United Kingdom
Email: m.j.inglis@lboro.ac.uk

Abstract

Much recent research attention has focused on understanding individual differences in the Approximate Number System, a cognitive system believed to underlie human mathematical competence. To date researchers have used four main indices of ANS acuity, and have typically assumed that they measure similar properties. Here we report a study which questions this assumption. We demonstrate that the Numerical Ratio Effect has poor test-retest reliability and that it does not relate to either Weber fractions or accuracy on nonsymbolic comparison tasks. Furthermore, we show that Weber fractions follow a strongly skewed distribution and that they have lower test-retest reliability than a simple accuracy measure. We conclude by arguing that in future researchers interested in indexing individual differences in ANS acuity should use accuracy figures, not Weber fractions or Numerical Ratio Effects.

INDEXING THE APPROXIMATE NUMBER SYSTEM

How do students develop their mathematical competence? In recent years there has been substantial interest in addressing this question by investigating individual differences in children and adults' abilities when performing basic arithmetic operations on nonsymbolic stimuli. Infants, children, adults and non-human animals are all capable of forming rapid nonsymbolic representations of the numerosity of arrays of dots and sequences of tones (e.g., Cordes, Gelman, Gallistel, & Whalen, 2001; Dehaene, 1997; Feigenson, Dehaene, & Spelke, 2004). The mechanism that underlies these representations has become known as the Approximate Number System (or ANS) and allows individuals to compare, add, and subtract sets of items, e.g., objects, dots, or tones (Barth, La Mont, Lipton, Dehaene, Kanwisher & Spelke, 2006; Meck & Church, 1983; Pica, Lemer, Izard, & Dehaene, 2004).

Some researchers have hypothesised that the ANS is the cognitive basis of all formal symbolic mathematics abilities; several sources of evidence support this view. First, the ANS is automatically activated in response to Arabic numerals in addition to nonsymbolic arrays (Moyer & Landauer, 1967). Second, prior to formal mathematical instruction children seem to be capable of using ANS mechanisms to perform approximate calculations with Arabic numerals despite being incapable of performing exact calculations (Gilmore, McCarthy, & Spelke, 2007). Third, measures of the precision of children's ANS representations – their so-called ANS acuity – have been found in some studies to predict their achievement on standardised school mathematics tests (e.g., De Smedt, Vershaffel, & Ghesquièrre, 2009; Halberda, Mazzocco, & Feigenson, 2008; Inglis, Attridge, Batchelor, & Gilmore, 2011; Libertus, Feigenson, & Halberda, 2011; Mazzocco, Feigenson, & Halberda, 2011a; Mundy & Gilmore, 2009). Fourth, it has been found in some studies that students with dyscalculia have lower ANS precision than typically achieving children, suggesting that an ANS deficit may be the cause of mathematical learning difficulties (Mazzocco, Feigenson & Halberda 2011b; Piazza, Faocertti, Trussardi, Berteletti, Conte, Lucangeli, Dehaene & Zorzi, 2010).

All these studies rely upon measuring an individual's ANS acuity: the accuracy with which they represent nonsymbolic numerosities. Typically this is achieved using the nonsymbolic comparison task. Participants are presented with two dot arrays n_1 and n_2, side by side or sequentially, and asked to judge which is the larger. After the presentation of many such pairs, one of four indices is typically calculated: accuracy, Weber fraction, numerical ratio effect (NRE) for accuracy or NRE for reaction time. These four indices are implicitly assumed to be measuring the same property: the acuity of an individual's ANS (e.g. Libertus, Feigenson, & Halberda, 2012; Price, Palmer, Battista, & Ansari, 2012). But, to date, little evidence has been presented for this suggestion. Our goal in this paper is to investigate the psychometric properties of, and interrelations between, these different indices. Before motivating our specific questions, we briefly discuss each of the four indices.

Several researchers have, when investigating ANS acuity, simply reported participants' accuracies: the proportion of trials they answered correctly (e.g., Fuhs & McNeil, 2013; Gilmore, Attridge, & Inglis, 2011; Kolkman, Kroesbergen, & Leseman, 2013; Lourenco, Bonny, Fernandez, & Rao, 2012; Nys, Ventura, Fernandes, Querido, Leybaert, & Content, 2013; Wei, Yuan, Chen & Zhou, 2012) or, less commonly, the number of trials they answered correctly in a given time (e.g., Nosworthy, Bugden, Archibald, Evans & Ansari, 2013).

The Weber fraction is an alternative approach to indexing an individual's ANS acuity (e.g. Bonny & Lourenco, 2013; Castronovo & Göbel, 2012; Halberda & Feigenson, 2008; Halberda et al., 2008; Halberda, Ly, Willmer, Naiman, & Germine, 2012; Inglis et al., 2011; Libertus et al., 2011, 2012; Lyons & Beilock, 2011; Mazzocco et al., 2011a; Piazza et al., 2010; Price et al., 2012; Sasanguie, Göbel, Moll, Smets, & Reynvoet, 2013). It makes the theoretical assumption that the ANS operates according to the Weber-Fechner law (e.g. Barth et al., 2006). Under this interpretation, when an individual observes an array of *n* dots, they form an internal representation which follows a normal distribution with mean *n* and standard deviation *wn*. Here *w* is the Weber fraction, which represents the precision of the individuals' representation. Those with *w*s closer to zero are more likely to form representations closer to the true value of the numerosity *n*. These assumptions imply that an individuals' expected accuracy on a given trial is a function of n_1, n_2 and *w*:

$$\mathrm{acc}(n_1, n_2; w) = \frac{1}{2} + \frac{1}{2}\mathrm{erf}\left(\frac{|n_1 - n_2|}{\sqrt{2}w\sqrt{n_1^2 + n_2^2}}\right)$$

. In practice, an individual's Weber fraction can be estimated by calculating the value of *w* which bests fits their behavioral data.

Figure 1 shows the relationship between the ratio of the two to-be-compared numerosities, and an individuals' expected accuracy for various values of *w*. As can be seen, as expected accuracy tends to 0.5, *w* asymptotically tends to infinity. It is therefore impossible for an individual to have an accuracy of under 0.5 under this model (to do so would require a *w* greater than infinity). The practical consequence of these considerations is that Weber fractions can only be calculated for participants whose responses follow the Weber-Fechner law, and consequently who score above 0.5 (cf. Libertus et al., 2011).

Finally, some researchers have adopted the Numerical Ratio Effect (NRE), or the closely related Numerical Distance Effect (NDE), to index ANS acuity (e.g., Bugden, Price, McLean & Ansari, 2012; Gilmore et al., 2011; Holloway & Ansari, 2009; Lonnemann, Linkersdorfer, Hasselhorn & Lindberg, 2011; Merkley & Ansari, 2010; Price et al., 2012; Sasanguie, Van den Bussche, & Reynvoet, 2012; Vanbinst, Ghesquière, & De Smedt, 2012). This effect observes that individuals are typically less accurate on, and slower to respond to, comparison trials where the n_1/n_2 ratio is close to 1. An individual's NRE can be obtained by calculating the slope of their ratio-accuracy (or ratio-RT) graph. Assuming that $n_1/n_2 < 1$, then an individual with a strongly negative NRE(accuracy) shows a substantial drop off in accuracy between easier trials (with ratios away from 1) and harder trials (those with ratios close to 1). Similarly, an individual with a strongly positive NRE(RT) shows a substantial slowing between easier and harder trials. As Figure 1 illustrates, the slope of an individuals' ratio-accuracy curve is predicted by their Weber fraction (the slope of the $w = 0.1$ curve is substantially steeper than the slope of the $w = 0.4$ curve, for example). Therefore, an individuals' NRE(accuracy) should, according to the standard model of the ANS, be strongly related to their Weber fraction (albeit non-linearly). It is less clear whether theory would predict a relationship between Weber fractions and NRE(RT)s, although many researchers have used NRE(RT) to index ANS acuity (e.g. Price et al., 2012).

The four different methods of indexing ANS acuity have, to a large extent, been assumed to unproblematically measure the same phenomenon (e.g., Libertus et al., 2012; Price et al., 2012). However, there are at least four reasons to doubt this belief.

First, calculations of the reliability of the different indices have been surprisingly low. Price et al. (2012) calculated immediate test-retest reliability figures for the NRE(RT) and Weber fraction on three variants of the nonsymbolic comparison task, finding reliability coefficients varying between $r = .4$ and .8; Maloney, Risko, Preston, Ansari & Fugelsang (2010) found that the immediate test-retest reliability of an NDE(accuracy) measure was in the same range, $r \approx .6$. Remarkably, Libertus et al (2012) found that the three month test-retest reliability of their measure of individuals' Weber fractions was not significantly different to zero.

Second, researchers have observed surprisingly low correlations between estimates of these indices obtained from different tasks. For example, Gilmore et al. (2011) found that estimates of Weber fraction obtained from a nonsymbolic comparison task did not correlate with similar indices derived from a nonsymbolic addition task, which is believed to be a closely related method of assessing ANS acuity (e.g. Barth et al., 2006).

Third, researchers have reported different relationships between their measures of individuals' ANS acuity and mathematical achievement. While some of these researchers have indexed ANS acuity using Weber fractions (e.g. Castronovo & Gobel, 2012; Halberda et al., 2008, 2012; Halberda & Feigenson, 2008; Inglis et al., 2011; Libertus et al., 2012; Lyons & Beilock, 2011; Piazza et al., 2010; Price et al., 2012; Sasanguie et al., 2012), others have used NREs (e.g., Bugden et al., 2012; Holloway & Ansari, 2009; Lourenco et al., 2012; Merkley & Ansari, 2010; Price et al., 2012), and others accuracy (e.g. Fuhs & McNeil, 2013; Nys et al., 2013; Wei et al., 2012). One account for why some of these researchers have found a relationship between ANS acuity and mathematics achievement, and others have not, is simply that their choice of index do not measure the same underlying phenomenon. For example, Mundy and Gilmore (2009) found a significant relationship between nonsymbolic comparison performance and mathematical achievement, but only when they indexed performance by accuracy rather than NDE.

Finally, to our knowledge, the only attempt to understand the relationship between different indices of ANS acuity suggests that the indices may measure different phenomena. Price et al. (2012) found extremely weak relationships between NDE(RT)s and Weber fractions. They found a significant (but weak, $R^2 = .11$) association between these two indices on a nonsymbolic comparison task where the stimuli were presented sequentially, and no significant associations on tasks where the stimuli were displayed concurrently.

To summarise, although much progress has been made towards understanding the ANS and its relationship with mathematical achievement, there is little agreement in the literature about how best to index an individual's ANS acuity. Further, there are reasons to suppose that at least some contradictory findings reported in the literature could be resolved by a careful study of the psychometric properties of different indices of the ANS. In this paper we take a step in this direction by asking four main questions. First, what distributions do the four commonly-used indices of ANS acuity (accuracy, Weber fraction, NRE(accuracy) and NRE(RT)) follow? Second, what are the relationships between these different indices? Third, what are the immediate and delayed test-retest reliabilities of the different indices? Finally, to what extent are accuracies and Weber fractions dependent on the problem sets from which they are derived? To answer these questions, we asked groups of adults and children to tackle a typical 80-trial nonsymbolic comparison task four times, twice in succession and then twice more in succession a week later.

1. Method

1.1 Participants.

Participants were 49 adults (ages 18-52, $M = 33.6$, 21 female), all staff or students working or studying at Loughborough University; and 56 children (ages 7-9, 25 female) recruited from a local school. The adults were paid £4 for participation, and the children were rewarded with stickers.

1.2 Materials and Procedure

The study took place during two sessions, scheduled exactly a week apart (within ±30 minutes). During each session participants were first familiarised with the nonsymbolic comparison task before being given 10 practice trials, and 80 experimental trials (Quarter 1). They were then asked to complete the same 80 experimental trials for a second time (Quarter 2). Participants repeated the same procedure in the second session. Consequently, by the end of the second session participants had completed the set of 80 trials four times (Quarters 1 to 4).

Each of the trials consisted of two dot arrays – one red, one blue – presented side by side on a laptop. Numerosities varied from 5-30, with each trial containing arrays that differed by one of four approximate ratios: 0.5, 0.6, 0.7 or 0.8. Participants were asked to select, as quickly and accurately as possible, which array was more numerous using the leftmost or rightmost buttons on a five-button response box. Stimuli were displayed until participants responded, or for 1000ms (to prevent counting). If participants had not responded within the time limit, a second screen, displaying a question mark, was shown until response. The side of the correct answer and its colour were counterbalanced across trials. To avoid participants relying upon continuous quantities associated with numerosity (e.g. dot size, envelope size), the stimuli were created using the method devised by Gebuis and Reynvoet (2011a). According to this method four sets of images were created, where: (1) envelope area and dot size are both positively correlated with the number of dots; (2) envelope area is positively correlated and dot size is negatively correlated with the number of dots; (3) envelope area is negatively correlated and dot size is positively correlated with the number of dots; and (4) envelope area and dot size are both negatively correlated with the number of dots. Participants were encouraged to take breaks every 20 trials, and while the children were given regular generic encouragement, no formal feedback about participants' accuracies was given.

2. Results and discussion

We structure our discussion of the results in four main sections. First, we report the distributions of the four indices under consideration. Second, we investigate their inter-relationships. Third, we explore the immediate and one-week test-retest reliabilities of the indices. Finally, we investigate the extent to which accuracies and Weber fractions depend on the problem sets from which they were derived.

2.1 Inclusion criteria and index construction

Participants were included in the analyses if they scored significantly above chance (on a sign test) on all four quarters of the experiment. A total of 21 participants (all children) failed to meet this criterion and were eliminated, leaving 84 participants in the final analysis.

For each participant we calculated, separately for each quarter of the experiment, their accuracy, their Weber fraction (by fitting their data to the model given above using maximum likelihood estimation), their NRE(accuracy) and their NRE(RT). Like earlier researchers (e.g., Libertus et al., 2011), we included *w* data from all participants for whom we were able to calculate Weber fractions, and so all 84 participants were represented in this dataset. Nine participants (all children) had outlier NRE(RT)s in one or more of the four quarters of the experiment (more than 3 SDs away from the mean). These participants were eliminated from analyses involving NRE(RT) but were retained in the other (non-RT) analyses. Descriptive statistics for the four indices under consideration are shown in Table 1.

2.2 Distributions of the indices

Figure 2 shows histograms of the indices for the children and adults separately. All the indices were roughly normally distributed, with the exception of the Weber fractions, which appeared to be strongly positively skewed (mean skewness statistics were 0.75 and 1.19 for adults and children respectively). We conducted Shapiro-Wilks tests of normality for each quarter of the experiment for adults and children separately. The distributions of Weber fractions significantly departed from normality for every one of the eight quarters, all ps < .003. In contrast, the accuracy and NRE(RT) distributions significantly departed from normality for only one of the eight quarters (children, quarter 1), and NRE(accuracy) never significantly departed from normality.

In sum, the accuracy and both NRE indices followed approximately normal distributions in both children and adults. However, the Weber fractions were not normally distributed for either children or adults. Instead they appeared to be strongly positively skewed. This finding is unsurprising in view of the function which relates accuracy and *w* (as discussed above, as accuracy tends to 0.5, *w* tends to infinity). However, it does have significant implications for how Weber fractions can be legitimately used by researchers interested in individual differences in ANS acuity. Specifically, most researchers who have used Weber fractions to index individual differences in ANS acuity have gone on to correlate them with, for example, measures of mathematical achievement. However, the Pearson correlation coefficient is not robust to violations of the assumption of normality (e.g. Bishara & Hittner, 2012; Kowalski, 1972), and Osborne (2010) suggested that correlational analyses should not be performed on data from distributions with a skew greater than 1.0 (which was the case in our child sample). This raises concerns about the appropriateness of analysing Weber fractions in this fashion without first performing an appropriate transformation (e.g. Osborne, 2008). In the general discussion section we discuss this issue further, and suggest that the skewedness of the distribution of Weber fractions can account for some of the results we report below, as well as several existing observations in the literature.

2.3 Inter-relations among the indices

Next we considered the extent to which the different indices were related to each other. Because we were interested in how the indices represented the same set of experimental data (rather than generalizing to the population of individuals), we used each quarter of the experiment (i.e. each complete set of 80 trials) as the unit of the analysis rather than each participant. Scatterplots showing how the different indices relate to each other are given in Figure 3. R^2 values for the relationships between the indices are given in Table 2. All these values reached significance, with the exception of the relationship between the NRE(accuracy) and NRE(RT), $R^2 = .01$, $p = .091$. Despite mostly showing significant relationships, the R^2 values were remarkably low: only the relationship between accuracy and

Weber fraction indicated that the two indices could reasonably be interpreted to be measuring the same construct: Cohen and Swedlik (2009), for example, suggested that any reliability figure below $r = .65$ ($R^2 = .42$) should normally be considered unacceptable. Furthermore, Figure 3 clearly indicates that the relationship between accuracies and Weber fractions is non-linear. Fitting a power law function, $y = ax^k$, rather than a straight line, gave a higher value of $R^2 = .86$. We also note that our data replicate Price et al.'s (2012) finding of a low relationship between NRE(RT) and Weber fractions. Whereas Price et al. found R^2s < .15, we found an R^2 of .14.

In sum, these analyses indicate that Weber fractions and accuracy figures can reasonably be interpreted to be measuring the same construct, presumably ANS acuity. But the same is not the case for NREs: neither the accuracy- or RT-based NREs were strongly related to accuracy or Weber fraction indices, and nor were they strongly related to each other. The lack of a strong relationship between NRE(accuracy)s and Weber fractions may have theoretical as well as practical significance. As shown in Figure 1, an individual's Weber fraction is, according to the standard model of the ANS, strongly, albeit non-linearly, related to the slope of their ratio-accuracy graph. We found no such relationship in our data: fitting a linear slope to the data resulted in an extremely low $R^2 = .16$, as did fitting a cubic, $R^2 = .19$, a curve which would be expected to better capture the theoretical non-linear relationship between NRE(accuracy)s and Weber fractions. This lack of a strong relationship may indicate that performance on the nonsymbolic comparison task is not governed by purely Weberian processes (cf. Fuhs & McNeil, 2013; Gebuis, Cohen Kadosh, de Haan, & Henik, 2009; Gebuis & Reynvoet, 2011b, 2012; Gilmore et al., 2013; Hurewitz, Gelman & Schnitzer, 2006; Nys & Content, 2012; Verguts & Fias, 2004; Zorzi & Butterworth, 1999). Alternatively, it may be that there is simply too much noise in behavioral data to observe this predicted theoretical relationship. That we found a strong relationship between Weber fractions and accuracy suggests that any noisiness which disrupts the NRE-Weber fraction relationship may be more present in the NRE measures than the Weber fraction. This suggestion is supported further by our test-retest analyses, reported below.

2.4 Test-retest reliability

We calculated four test-retest reliability coefficients for each index. Two provided estimates of the indices' immediate test-retest reliabilities: the correlations between Quarters 1 and 2, and between Quarters 3 and 4. Two more provided estimates of the indices' one week test-retest reliabilities: the correlations between Quarters 1 and 3, and between Quarters 2 and 4.

The resulting four scatterplots for each of the four indices are shown in Figures 4 – 7. Means of the two estimates for each indices' immediate and one-week test-retest reliability coefficients are given in Table 3 (top panel). Of the four indices, accuracy showed the highest reliability (which, for adults, was in Cohen and Swerdlik's (2009) acceptable range of $r \geq .65$), followed by Weber fraction (although all these reliability figures remained in the unacceptable range). Both the NRE indices showed extremely poor test-retest reliability, and in the child sample the figures were close to zero.

To investigate the extent to which increasing the length of a nonsymbolic numerical comparison task would influence its one-week test-retest reliability, we combined Quarters 1 and 2, and 3 and 4, into two halves, each of 160 trials and recalculated indices for accuracy, Weber fraction, NRE(accuracy) and NRE(RT). These figures are shown in Table 3 (bottom panel). This substantially increased the one-week test-retest reliability for accuracy in adults (from .65 to .79), but surprisingly had little effect on the accuracy figure for children, or on the figures for Weber fraction (in adults or children). As before, the NRE reliability figures for children remained close to zero.

When reporting their three-month test-retest analysis of Weber fractions, Libertus et al. (2012) noticed that some of their participants showed greater between session variability than others. In particular, they found that the size of this variability was greater for participants with high Time 1 Weber fractions (i.e. lower ANS acuity), compared to those with low Weber fractions at Time 1. To investigate this observation, we followed Libertus et al.'s approach by calculating each individuals' absolute change in Weber fraction from Quarter 1 to Quarter 3 (repeating this analysis for Quarter 2 to Quarter 4 changes yielded essentially identical results). These were significantly correlated with Weber fractions from Quarter 1, $r = .372$, $p = .001$, replicating Libertus et al.'s finding.

However, we do not believe that this finding indicates, as suggested by Libertus et al. (2012), that those participants with low ANS acuity (i.e. high Weber fractions) show particularly labile ANS acuities. Instead we suggest that this correlation is merely the result of the positive skewness of the distribution of Weber fractions, coupled with regression to the mean. Those participants with high Weber fractions were relatively further from the mean of the distribution than those with low Weber fractions (because of the skewed distribution), so they would be expected to regress further in subsequent measurements. To test this alternative account, we conducted two further analyses. First, we repeated Libertus et al.'s analysis with accuracy figures rather than Weber fractions (that is to say that we correlated participants' Quarter 1 accuracies with their absolute change in accuracy from Quarter 1 to Quarter 3). This revealed no significant correlation, $r = -.089$, $p = .422$, which would be surprising if participants with low ANS acuities had particularly variable performance. Second, we tested whether the relationship observed by Libertus et al. could also be observed in reverse. A signature of regression to the mean effects is that they are time reversible, that is to say that both large Time 1 and large Time 2 scores should 'cause' larger absolute Time 1 to Time 2 differences. We found exactly this pattern: Quarter 3 Weber fractions were significantly related to participants' absolute Quarters 1 to 3 change scores, $r = .779$, $p < .001$.

The positive skew of the distribution of Weber fractions may also account for the difference in reliability observed between accuracies and Weber fractions. Recall that, for adults at least, the test-retest reliabilities observed for Weber fractions were consistently within Cohen and Swerdlik's (2009) unacceptable range ($r < .65$), whereas the equivalent figures for accuracy were consistently above it. Because of the power law relationship between accuracies and Weber fractions, small changes in accuracy have different expected effects on the Weber fraction depending on whether the individual is a high or low performer. For example, the curve of best fit for the accuracy-Weber fraction graph shown in Figure 3 is given by $w = 0.13a^{-4.1}$ (where w is the Weber fraction and a is accuracy). Thus, if an individual moves from $a = .95$ to $a = .94$, we would expect their Weber fraction to change by only 0.007. But if a different individual's accuracy changed from $a = .55$ to $a = .54$, their Weber fraction would be expected to change by 0.12, a figure nearly 17 times greater. Of course, this different effect would be in some sense desirable if participants were following the Weber-Fechner law (a change in accuracy from .54 to .55 would be more significant than a change from .94 to .95 in terms of the precision of their ANS acuity), however if dot comparison performance is not entirely Weberian (as suggested by, for example, Fuhs & McNeil, 2013; Gebuis et al., 2009; Gilmore et al., 2013; Hurewitz et al., 2006; Nys & Content, 2012; Verguts & Fias, 2004; Zorzi & Butterworth, 1999), then this property of the Weber fraction may be less desirable. In short, it may be that this exaggeration of small changes in behavior at the lower end of performance accounts for the relatively poor test-retest reliability observed in Weber fractions.

In sum, we found widely varying levels of test-retest reliability for the different indices used to investigate individual differences in ANS acuity. Both the NRE indices had extremely low test-retest reliability, especially in children. The most reliable index was accuracy. The test-retest reliabilities of the accuracy indices for adults for both the 80- and 160-item tests were within the range considered acceptable by psychometricians (e.g. Cohen & Swerdlik, 2009), but the equivalent figures for children, were somewhat lower than this. Weber fractions proved to have reliability coefficients some distance away from the acceptable range for both the 80- and 160-item test. It is possible that Weber fractions' skewed distribution may account for their comparatively low reliability coefficients.

2.4 Ratio Invariance of Accuracies and Weber fractions.

One clear difference between the accuracy and Weber fraction indices is that accuracies are strongly related to the ratios of the to-be-compared numerosities on a nonsymbolic comparison task. Ratios close to 1 are more difficult than those further away from 1, therefore one can manipulate a participant's accuracy by choosing harder or easier ratios. Is the same true of Weber fractions, or are they largely independent of the presented ratios? If the latter were true, one could more easily compare Weber fractions derived from different problem sets than one could accuracies.

To investigate this issue we split our problems (across all quarters) into two sets. Set A consisted of 75% of the trials with ratios 0.5 and 0.6, and 25% of the trials with ratios 0.7 and 0.8, and Set B was formed of the remaining trials. Thus we had two sets of problems, one with an average ratio of 0.6 and one with an average ratio 0.7, but both with a range of 0.5 to 0.8. We then calculated each participants' accuracy and Weber fraction for the two sets of problems separately. To compare the between-problem-set variance of accuracies and Weber fractions, we standardised each measure (across both problem sets), and calculated, for each participant, their absolute difference. Both individuals' accuracies, $r = .786$, $p < .001$, and their Weber fractions, $r = .752$, $p < .001$, were strongly correlated across problem sets (these figures are in some sense measures of the indices' internal reliabilities). As expected, participants' accuracies differed by a mean of 1.24 standard deviations between Set A and Set B. We also found a smaller, but still substantial, mean difference of 0.28 standard deviations between the Weber fractions derived from Set A and those derived from Set B. For both adults and children, the Set A – Set B differences in both accuracies and Weber fractions were signficant, all *p*s < .001

These figures suggest that both accuracies and Weber fraction seem to depend, in part, on the ratios of the problem sets from which they are derived. Although this effect was substantially smaller for Weber fraction than it was for accuracies, it nevertheless seems reasonable to conclude that both indices are influenced by researchers' choices of ratios.

3. General Discussion

In recent years there has been increasing interest in individual differences in ANS acuity, but researchers have adopted several different methodological approaches in their investigations. Here we considered four commonly used indices of ANS acuity: accuracy, Weber fraction and two variants of the NRE. We considered the psychometric properties of these indices, and how they were related, reporting four main findings. First, that accuracy and NREs for both accuracy and RT are distributed approximately normally, but that Weber fractions are distributed with a strongly positive skew. Second, that accuracy and Weber fractions are strongly related, $R^2 = .86$, suggesting that they are measuring the same underlying construct, but that accuracy- and RT-based NREs are not related to each other, and neither are they related to accuracies or Weber fractions. Third, we reported the immediate and one-week test-retest reliabilities of the four measures, finding that both the NRE indices had reliabilities close to zero in the child sample, and that accuracies had stronger test-retest reliability than Weber fractions in both the immediate and one-week analyses, for both adults and children. Finally, we found that both accuracies and Weber fractions depend, in part, upon the ratios of the problem sets from which they are derived.

In our view, these findings clearly indicate that researchers should no longer use NRE measures when trying to index ANS acuity, especially when investigating young populations. NRE indices do not relate to other indices of ANS acuity, and have extremely low test-retest reliability in child samples. Furthermore, it may be that the lack of the predicted relationship between NRE(accuracy) and Weber fraction has some theoretical significance, implying that there are important processes involved in the nonsymbolic comparison task which do not follow the Weber-Fechner law (as suggested by, for example, Fuhs & McNeil, 2013; Gebuis, Cohen Kadosh, de Haan, & Henik, 2009; Gebuis & Reynvoet, 2011b, 2012; Gilmore et al., 2013; Hurewitz, Gelman & Schnitzer, 2006; Nys & Content, 2012; Verguts & Fias, 2004; Zorzi & Butterworth, 1999).

A second implication of our findings is that an individual's accuracy seems to be a substantially superior measure of ANS acuity to their Weber fraction, especially when a large number of trials are used. While these two indices are strongly related, accuracy shows higher test-retest reliability and follows a normal distribution. In contrast, the distribution of Weber fractions in our study was strongly positively skewed, and we suggested that this skew was responsible for the relatively low test-retest reliability we observed. Furthermore, we have suggested that the skewness of the Weber fraction distribution is responsible for the relationship, observed by Libertus et al. (2012), between an individual's Weber fraction and its subsequent change. Because the distribution is skewed, we would expect asymmetric regression to the mean effects. A further disadvantage of the Weber fraction is that as accuracies decrease to 0.5, they asymptotically tend to infinity. In fact, it would require a Weber fraction greater than infinity to successfully model individuals' who score less than 0.5. While such a score would appear impossible if the nonsymbolic comparison task involves only Weber-Fechner processes (a claim which is disputed by some researchers), some participants do indeed perform in this range: 10 of the 420 experimental quarters in our study resulted in accuracies below 0.5, and in Libertus et al.'s (2011) study, over 10% of participants fell into this category.

Given their worse reliability, is there any reason for researchers to continue to use Weber fractions to index the ANS ahead of simple accuracy scores? One justification would be because the Weber fraction is a theoretically-motivated measure which derives from the well-established psychophysical Weber-Fechner law. However, as we have noted, many researchers have questioned whether performance on dot comparison tasks is entirely Weberian. Some have highlighted the role that inhibitory control plays in visual dot comparison tasks (e.g. Fuhs & McNeil, 2013; Gebuis, Cohen Kadosh, de Haan, & Henik, 2009; Gilmore et al., 2013; Hurewitz, Gelman & Schnitzer, 2006; Nys & Content, 2012) and others have proposed alternative accounts entirely (e.g. Gebuis & Reynvoet, 2011b, 2012; Verguts & Fias, 2004; Zorzi & Butterworth, 1999). Given that there is no consensus on the processes by which nonsymbolic comparison takes place, relying upon a measure which is contingent on a particular theory could be seen as being somewhat premature.

Another justification for preferring Weber fractions could be to argue that they are superior to accuracy because they are a universal measure of ANS acuity which allow researchers to better compare performance between tasks. In contrast, accuracies clearly vary between experiments depending on the particular paradigm adopted, and numerosities involved. Indeed, some researchers have implicitly taken this position. For example, Piazza (2010) plotted Weber fractions derived from seven different sources on one graph, where each source had used substantially different experimental procedures. These differences can be categorized on several dimensions, including presentation style (habituation versus comparison), numerosity (ranges of 4-16, 5-16, 12-40 and up to 80) and stimuli duration (time limited or until response). Clearly this approach requires the assumption that Weber fractions derived from substantially different tasks are comparable. In a similar vein, some researchers have analyzed together data from participants who were asked to tackle the nonsymbolic comparison task with different stimuli durations (e.g., Halberda & Feigenson, 2008; Mazzocco et al. 2011a).

Both these approaches implicitly assume that the Weber fractions derived from these various methods are comparable, but there are at least three reasons to question this assumption. First, as we have shown in here, Weber fractions do appear to be, in part, influenced by the ratios of the trials from which they are derived. Second, Price et al. (2012) found that Weber fractions derived from extremely similar versions of the nonsymbolic comparison task were not strongly related. They gave participants nonsymbolic comparison tasks where the stimuli were presented either concurrently separately, concurrently overlapping or sequentially, finding low *r*s of .39, .50 and .68 (see also Gilmore et al., 2011). Third, Inglis and Gilmore (2013) found that Weber fractions derived from a nonsymbolic comparison task systematically varied with the duration for which numerical stimuli were displayed. These findings appear to suggest that it is unreasonable to assume that Weber fractions derived from studies with different stimuli and experimental designs are comparable.

In sum, we believe that it is unreasonable to assume that Weber fractions are paradigm independent. To be clear, we do not suggest that these issues effect Weber fractions but not accuracies: they effect *all* measures of ANS acuity. Specifically, both accuracy and Weber fraction indices can only be interpreted in relation to other figures derived from the same experimental design. Given this, and given the superior psychometric properties of simple accuracy figures, we suggest that the best way of indexing the acuity of an individual’s ANS is simply to report their accuracy on a large number of nonsymbolic comparison trials.

References

Aiken, L. R. (1994). *Psychological testing and assessment*. (8th ed.). Massachusetts: Simon & Schuster, Inc.

Barth, H., La Mont, K., Lipton, J., Dehaene, S., Kanwisher, N., & Spelke, E. (2006). Nonsymbolic arithmetic in adults and young children. *Cognition, 98*, 199-222.

Bishara, A., J., & Hittner, J. B. (2012). Testing the significance of a correlation with nonnormal data: comparison of Pearson, Spearman, transformation, and resampling approaches. *Psychological Methods, 17,* 399-417.

Bonny, J.W. & Lourenco, S. F. (2013). The approximate number system and its relation to early math achievement: Evidence from the preschool years. *Journal of Experimental Child Psychology, 114,* 375-388.

Bugden, S., Price, G. R., McLean, D. A., & Ansari, D. (2012). The role of the left intraparietal sulcus in the relationship between symbolic number processing and children's arithmetic competence. *Developmental Cognitive Neuroscience, 2,* 448-457.

Castronovo, J. & Göbel, S.M. (2012). Impact of high mathematics education on the number sense. *PLOS ONE 7*, e33832.

Cohen, R. J. & Swerdlik, M. (2009). *Psychological Testing and Assessment: An Introduction to Tests and Measurement* (7th edition). New York, NY: McGraw-Hill.

Cordes, S., Gelman, R., Gallistel, C.R., & Whalen, J. (2001). Variability signatures distinguish verbal from nonverbal counting for both large and small numbers. *Psychonomic Bulletin and Review, 8*, 698-707.

Dehaene, S. (1997). *The number sense*. Oxford, UK: Oxford University Press.

De Smedt, B., Verschaffel, L., & Ghesquière, P. (2009). The predictive value of numerical magnitude comparison for individual differences in mathematics achievement. *Journal of Experimental Child Psychology, 103*, 469–479.

Fuhs, M. W., & McNeil, N. M. (2013). ANS acuity and mathematics ability in preschoolers from low‐income homes: contributions of inhibitory control. *Developmental science*, *16*, 136-148.

Feigenson, L., Dehaene, S., & Spelke, E. S. (2004). Core systems of number. *Trends in Cognitive Sciences, 8*, 307–314.

Gebuis, T., Cohen Kadosh, R., de Haan, E., & Henik, A. (2009). Automatic quantity processing in 5-year olds and adults. *Cognitive Processing*, *10*, 133-142.

Gebuis, T. & Reynvoet, B. (2011a). Generating non-symbolic number stimuli. *Behavior Research Methods, 43*, 981-986.

Gebuis, T. & Reynvoet, B. (2011b). The interplay between visual cues and non-symbolic number. *Journal of Experimental Psychology: General, 141*, 642-648.

Gebuis, T. & Reynvoet, B. (2012). Continuous visual properties explain neural responses to non-symbolic number. *Psychophysiology, 49,* 1481-1491.

Gilmore, C., Attridge, N., Clayton, S., Cragg, L., Johnson, S., Marlow, N., ... & Inglis, M. (2013). Individual Differences in Inhibitory Control, Not Non-Verbal Number Acuity, Correlate with Mathematics Achievement. *PLOS ONE*, *8*(6), e67374.

Gilmore, C.K., McCarthy, S., & Spelke, E.S. (2007). Symbolic arithmetic knowledge without instruction. *Nature*, 447, 589-591.

Gilmore, C.K., Attridge, N., & Inglis, M. (2011). Measuring the approximate number system. *Quarterly Journal of Experimental Psychology.*

Halberda, J., Ly, R., Willmer, J., Naiman, D., & Germine, L. (2012). Number sense across the lifespan as revealed by a massive internet-based sample. *Proceedings of the National Academy of Sciences, 109*, 11116-11120.

Halberda, J., Mazzocco, M.M., & Feigenson, L. (2008). Individual differences in non-verbal number acuity correlate with maths achievement. *Nature, 455,* 665-668.

Halberda, J. & Feigenson, L. (2008). Developmental change in the acuity of the "number sense": The Approximate Number System in 3-, 4-, 5-, and 6-year-olds and adults. *Developmental Psychology, 44,* 1457-1465.

Hurewitz, F., Gelman, R., & Schnitzer, B. (2006). Sometimes area counts more than number. *Proceedings of the National Academy of Sciences*, *103*, 19599-19604.

Holloway, I.D. & Ansari, D. (2009). Mapping numerical magnitudes onto symbols: The numerical distance effect and individual differences in children's mathematics achievement. *Journal of Experimental Child Psychology, 103,* 17-29.

Inglis, M., Attridge, N., Batchelor, S. & Gilmore, C. (2011). Non-verbal number acuity correlates with symbolic mathematics achievement: But only in children. *Psychonomic Bulletin & Review, 18,* 1222-1229.

Inglis, M., & Gilmore, C. (2013). Sampling from the Mental Number Line: How are Approximate Number System Representations Formed? *Cognition, 129,* 63-69.

Kolkman, M. E., Kroesbergen, E. H., & Leseman, P. P. M. (2013). Early numerical development and the role of non-symbolic and symbolic skills. *Learning and Instruction, 25,* 95-103.

Kowalski, C. J. (1972). On the effects of non-normality on the distribution of the sample product-moment correlation coefficient. *Journal of the Royal Statistical Society. Series C, 21,* 1-12.

Libertus, M. E., Feigenson, L., & Halberda, J. (2011). Preschool acuity of the approximate number system correlates with school math ability. *Developmental Science, 14*, 1292-1300.

Libertus, M. E., Odic, D., & Halberda, J. (2012). Intuitive sense of number correlates with scores on college-entrance examination. *Acta Psychologica, 141,* 373-379.

Lonnemann, J., Linkersdorfer, J., Hasselhorn, M., & Lindberg, S. (2011). Symbolic and non-symbolic distance effects in children and their connection with arithmetic skills. *Journal of Neurolinguistics, 24,* 583-591.

Lourenco, S. F., Bonny, J. W., Fernandez, E. P., Rao, S. (2012). Nonsymbolic number and cumulative area representations contribute shared and unique variance to symbolic math competence. *PNAS, 109,* 18737-18742.

Lyons, I. M. & Beilock, S. L. (2011). Numerical ordering ability mediates the relation between number-sense and arithmetic competence. *Cognition, 121*, 256-261.

Maloney, E.A., Risko, E.F., Preston, F., Ansari, D., & Fugelsang, J. (2010). Challenging the reliability and validity of cognitive measures: The case of the numerical distance effect. *Acta Pyschologica, 134*, 154-161.

Mazzocco, M. M. M., Feigenson, L., & Halberda, J. (2011a). Preschoolers' precision of the approximate number system predicts later school mathematics performance. *PLOS ONE*, *6*(9), e23749.

Mazzocco, M. M. M., Feigenson, L., & Halberda, J. (2011b). Impaired Acuity of the Approximate Number System Underlies Mathematical Learning Disability (Dyscalculia). *Child Development*, *82*, 1224–1237.

Meck, W. H., & Church, R. M. (1983). A mode control model of counting and timing processes. *Journal of Experimental Psychology: Animal Behavior Processes, 9,* 320-334.

Merkley, R., & Ansari, D. (2010). Using eye tracking to study numerical cognition: the case of the ratio effect. *Experimental Brain Research, 206*, 455-60.

Moyer, R.S. & Landauer, T.K. (1967). Time required for judgements of numerical inequality. *Nature, 215*, 1519-1520.

Mundy, E. & Gilmore, C.K. (2009). Children's mapping between symbolic and nonsymbolic representations of number. *Journal of Experimental Child Psychology, 103*, 490-502.

Nosworthy, N., Bugden, S., Archibald, L., Evans, B., & Ansari, D. (2013). A Two-Minute Paper-and-Pencil Test of Symbolic and Nonsymbolic Numerical Magnitude Processing Explains Variability in Primary School Children's Arithmetic Competence. *PLOS ONE*, *8*(7), e67918.

Nys, J., & Content, A. (2012). Judgement of discrete and continuous quantity in adults: Number counts!. *The Quarterly Journal of Experimental Psychology*, *65*, 675-690.

Nys, J. Ventura, P., Fernandes, T., Querido, L., Leybaert, J., & Content, A. (2013). Does math education modify the approximate number system? A comparison of schooled and unschooled adults. *Trends in Neuroscience and Education, 2,* 13-22.

Osborne, J. W. (2008). Best practices in data transformation: The overlooked effect of minimum values. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 197–204). Thousand Oaks, CA: Sage.

Osborne, J. W. (2010). Correlation and other measures of association. In G. R. Hancock & R. O. Mueller (Eds.), *The Reviewer's Guide to Quantitative Methods in the Social Sciences* (pp. 55-70). New York, NY: Routledge.

Piazza, M., Facoetti, A., Trussardi, A. N., Berteletti, I., Conte, S., Lucangeli, D., Dehaene, S. & Zorzi, M. (2010). Developmental trajectory of number acuity reveals a severe impairment in developmental dyscalculia. *Cognition, 116,* 33-41.

Piazza, M. (2010). Neurocognitive start-up tools for symbolic number representations. *Trends in Cognitive Sciences, 14,* 542-551.

Pica, P., Lemer, C., Izard, V., & Dehaene, S. (2004). Exact and approximate arithmetic in an Amazonian indigene group. *Science*, *306*, 499-503.

Price, G. R., Palmer, D., Battista, C., & Ansari, D. (2012). Nonsymbolic numerical magnitude comparison: Reliability and validity of different task variants and outcome measures, and their relationship to arithmetic achievement in adults. *Acta Psychologica, 140*, 50-57.

Sasanguie, D., Gobel, S., Moll, K., Smets, K., & Reynvoet, B. (2013). Approximate number sense, symbolic number processing, or number–space mappings: What underlies mathematics achievement? *Journal of Experimental Child Psychology, 114*, 418-431.

Sasanguie, D., Van den Bussche, E., Reynvoet, B. (2012). Predictors for mathematics achievement? Evidence from a longitudinal study. *Mind Brain and Education, 6*, 119-128.

Vanbinst, K., Ghesquière, P., & De Smedt, B. (2012). Numerical Magnitude Representations and Individual Differences in Children's Arithmetic Strategy Use. *Mind, Brain and Education, 6,* 129-136.

Verguts, T., & Fias, W. (2004). Representation of number in animals and humans: a neural model. *Journal of Cognitive Neuroscience*, *16*, 1493-1504.

Wei, W., Yuan, H., Chen, C., Zhou, X. (2011). Cognitive correlates of performance in advanced mathematics. *British Journal of Educational Psychology, 82,* 157-181.

Zorzi, M. & Butterworth, B. (1999). A computational model of number comparison. In M. Hahn & S. C. Stoness (Eds.), *Proceedings of the 21st Annual Conference of the Cognitive Science Society* (pp. 772-777). Mahwah, NJ: Lawrence Erlbaum.

Author Note

This research was supported by a British Academy Postdoctoral Fellowship (C.G.), and a Royal Society Worshipful Company of Actuaries Educational Research Fellowship (M.I.).

Table Captions

Table 1. Descriptive statistics, averaged across each of the four quarters of the experiment, for each of the four indices under consideration.

Table 2. Inter-relations between the four different indices, given as R^2 values derived from Pearson correlation coefficients. All figures are significantly different to zero, except for the relationship between NRE(accuracy) and NRE(RT). Note that the relationship between accuracy and Weber fraction is non-linear, so the figure in the Table is an underestimate of the true relationship: fitting a power law function ($y = ax^k$) to these data yielded an R^2 value of .86.

Table 3. Immediate and one-week test-retest reliability coefficients for the four indices, separately for adults and children. The 80-trial figures are averaged over two measurements (the immediate figures are the mean of the correlations between Quarters 1 and 2, and 3 and 4, and the one-week figures are the mean of the correlations between Quarters 1 and 3, and 2 and 4). **p*s < .05, ***p*s < .01, ****p*s < .001.

Table 1. Descriptive statistics, averaged across each of the four quarters of the experiment, for each of the four indices under consideration.

	Mean	SD	Skewness	Kurtosis
		Children		
Accuracy	0.73	0.06	-0.15	-0.51
Weber fraction	0.51	0.18	1.19	1.62
NRE(accuracy)	-0.61	0.41	0.00	0.23
NRE(RT)	62	418	-0.19	0.93
		Adults		
Accuracy	0.85	0.06	-0.52	-0.41
Weber fraction	0.25	0.08	0.75	0.02
NRE(accuracy)	-0.89	0.33	-0.27	-0.05
NRE(RT)	427	258	-0.07	0.78

Table 2. Inter-relations between the four different indices, given as R^2 values derived from Pearson correlation coefficients. All figures are significantly different to zero, except for the relationship between NRE(accuracy) and NRE(RT). Note that the relationship between accuracy and Weber fraction is non-linear, so the figure in the Table is an underestimate of the true relationship: fitting a power law function ($y = ax^k$) to these data yielded an R^2 value of .86.

	Accuracy	Weber Fraction	NRE(accuracy)	NRE(RT)
Accuracy	-	.79	.02	.19
Weber Fraction	.79	-	.16	.14
NRE(accuracy)	.02	.16	-	.01
NRE(RT)	.19	.14	.01	-

Table 3. Immediate and one-week test-retest reliability coefficients for the four indices, separately for adults and children. The 80-trial figures are averaged over two measurements (the immediate figures are the mean of the correlations between Quarters 1 and 2, and 3 and 4, and the one-week figures are the mean of the correlations between Quarters 1 and 3, and 2 and 4). **p*s < .05, ***p*s < .01, ****p*s < .001.

	Immediate		One-week	
	Adults	Children	Adults	Children
Accuracy (80 trials)	.68***	.57**	.65***	.47**
Weber fraction (80 trials)	.55**	.50*	.60***	.41*
NRE(accuracy) (80 trials)	.28	-.02	.27	-.13
NRE(RT) (80 trials)	.27	.21	.27	-.07
Accuracy (160 trials)			.79***	.52**
Weber fraction (160 trials)			.59***	.47**
NRE(accuracy) (160 trials)			.52***	-.05
NRE(RT) (160 trials)			.53***	.11

Figure Captions

Figure 1. Predicted accuracy as a function of the n_1/n_2 ratio, for various values of w.

Figure 2. Histograms showing the distributions of the four indices, separately for adults and children, combined across all quarters of the experiment.

Figure 3. Scatterplots showing how each of the indices relate to each other. R^2 figures are derived from Pearson correlations coefficients, other than for the accuracy – Weber fraction plot, which is derived from a power law function.

Figure 4. The test-retest reliability of the accuracy index.

Figure 5. The test-retest reliability of the Weber fraction index.

Figure 6. The test-retest reliability of the NRE(accuracy) index.

Figure 7. The test-retest reliability of the NRE(RT) index.

Figure 1. Predicted accuracy as a function of the n_1/n_2 ratio, for various values of *w*.

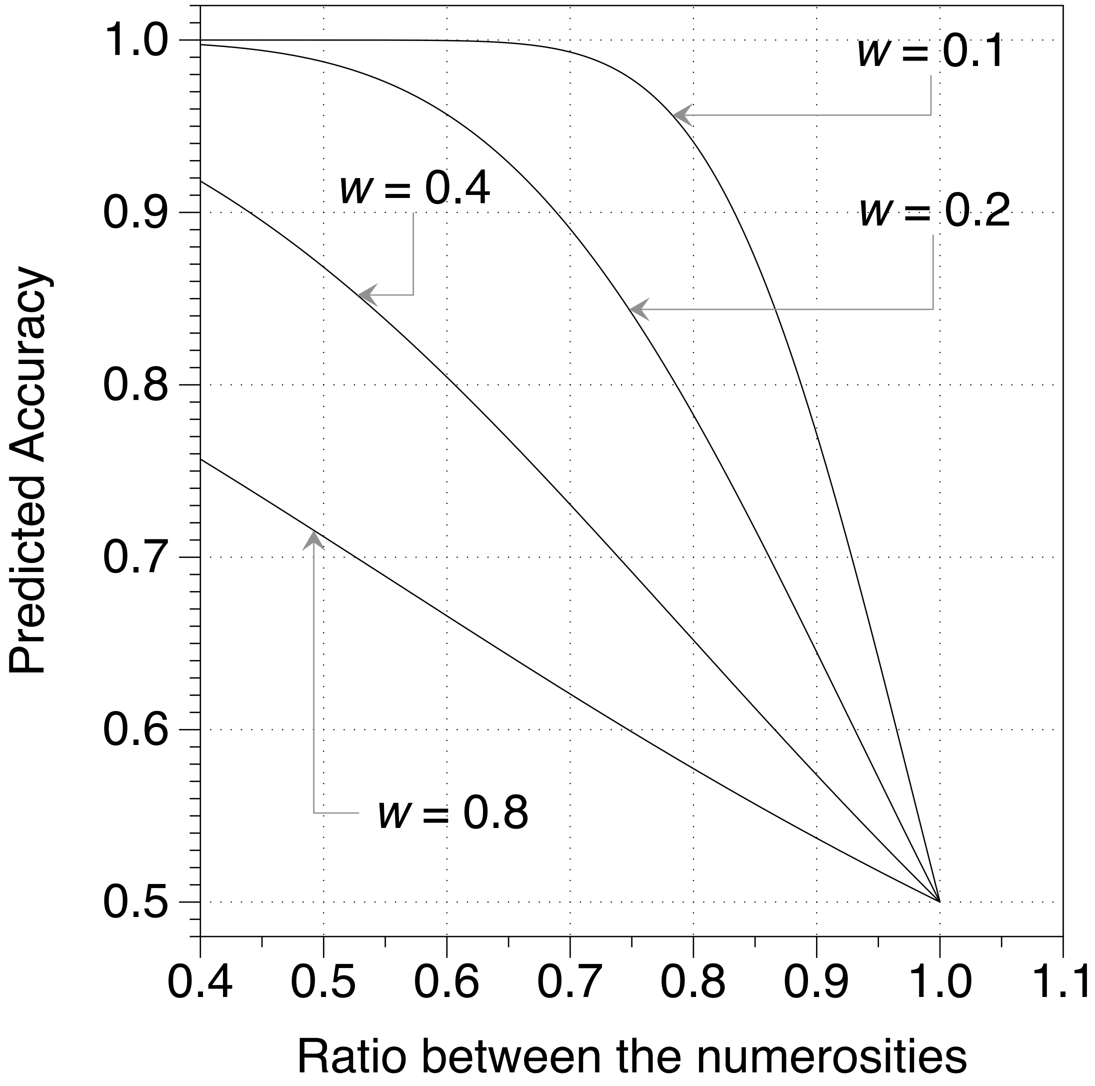


Figure 2. Histograms showing the distributions of the four indices, separately for adults and children, combined across all quarters of the experiment.

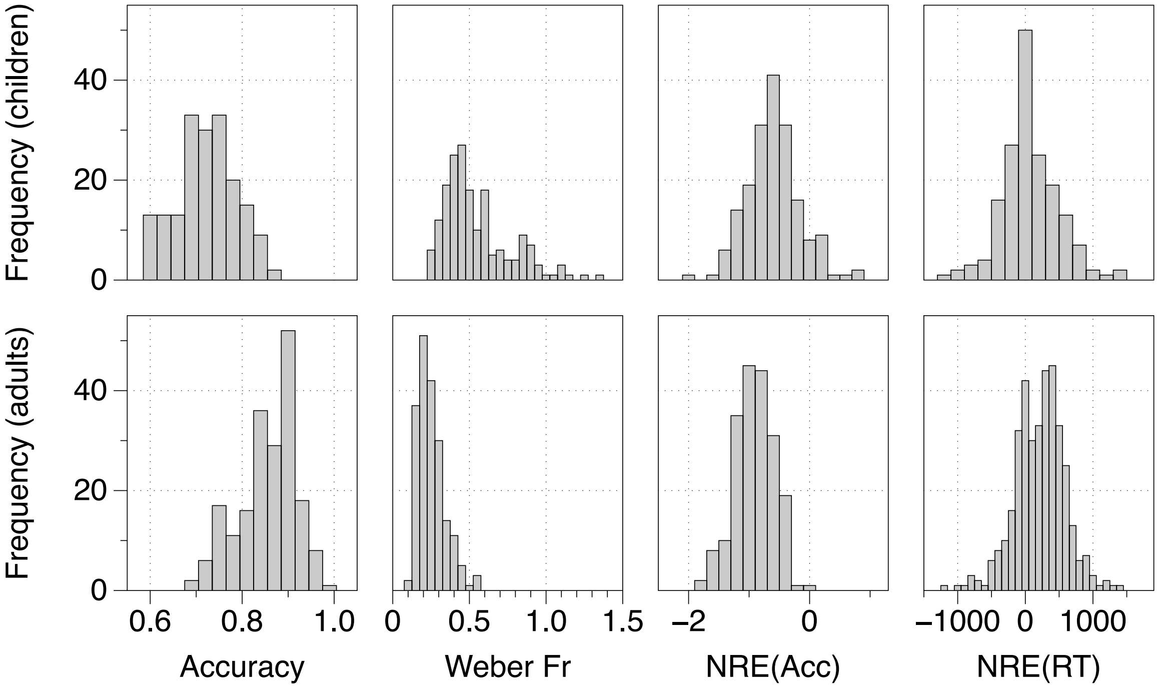


Figure 3. Scatterplots showing how each of the indices relate to each other. R^2 figures are derived from Pearson correlations coefficients, other than for the accuracy – Weber fraction plot, which is derived from a power law function. All R^2 are significant other than the NRE(RT) – NRE(acc) relationship.

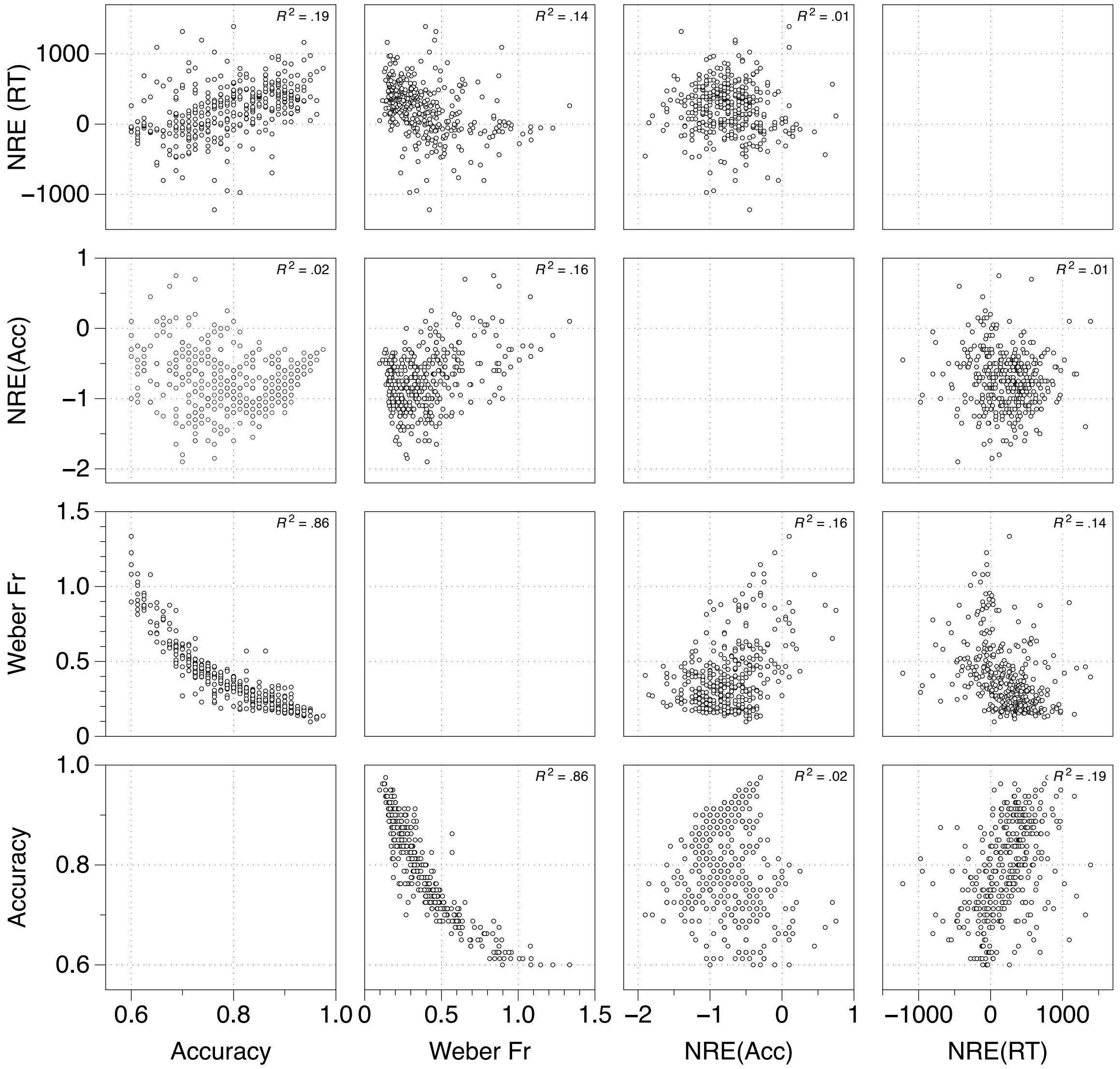


Figure 4. The test-retest reliability of the accuracy index.

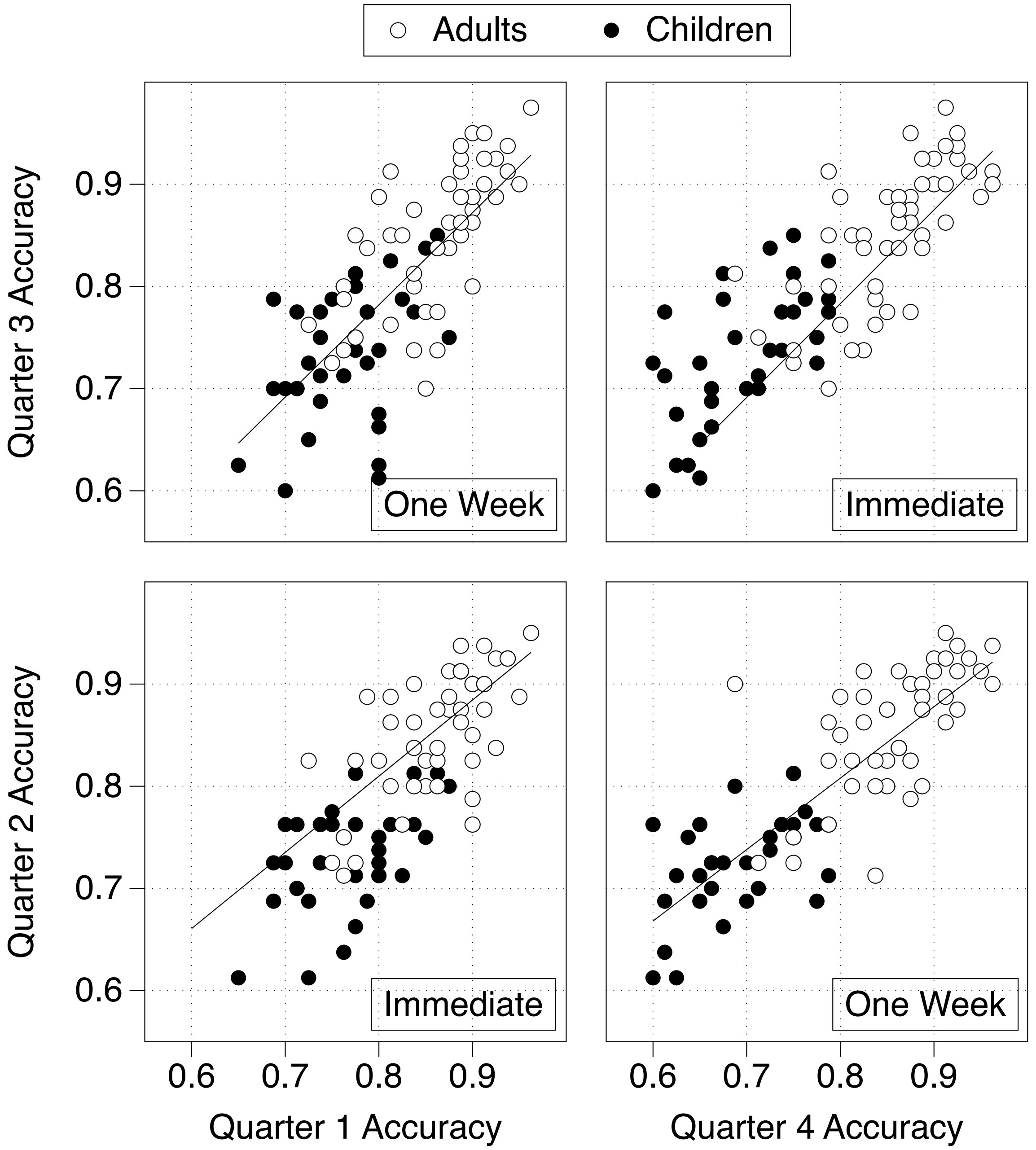


Figure 5. The test-retest reliability of the Weber fraction index.

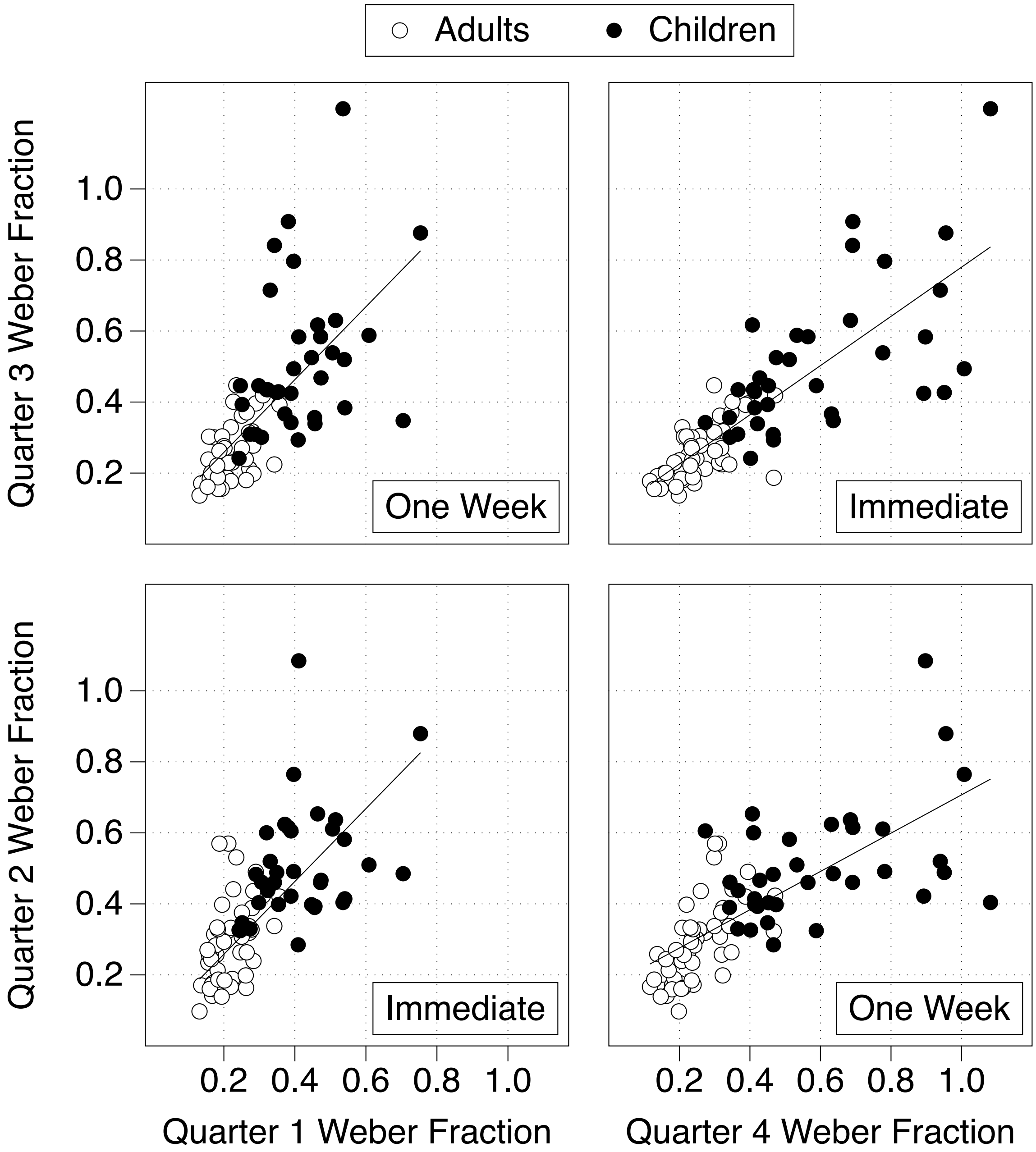


Figure 6. The test-retest reliability of the NRE(accuracy) index.

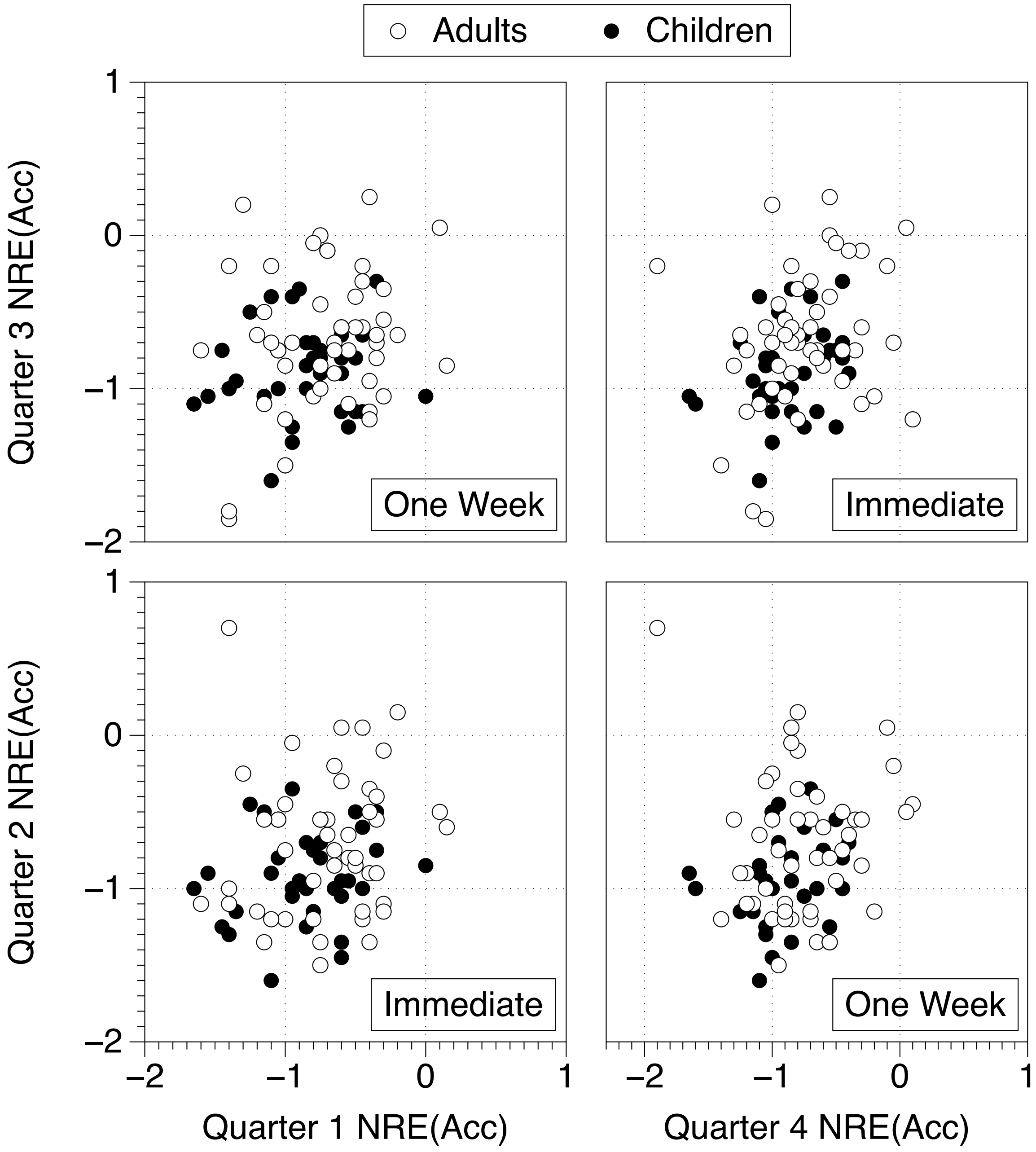
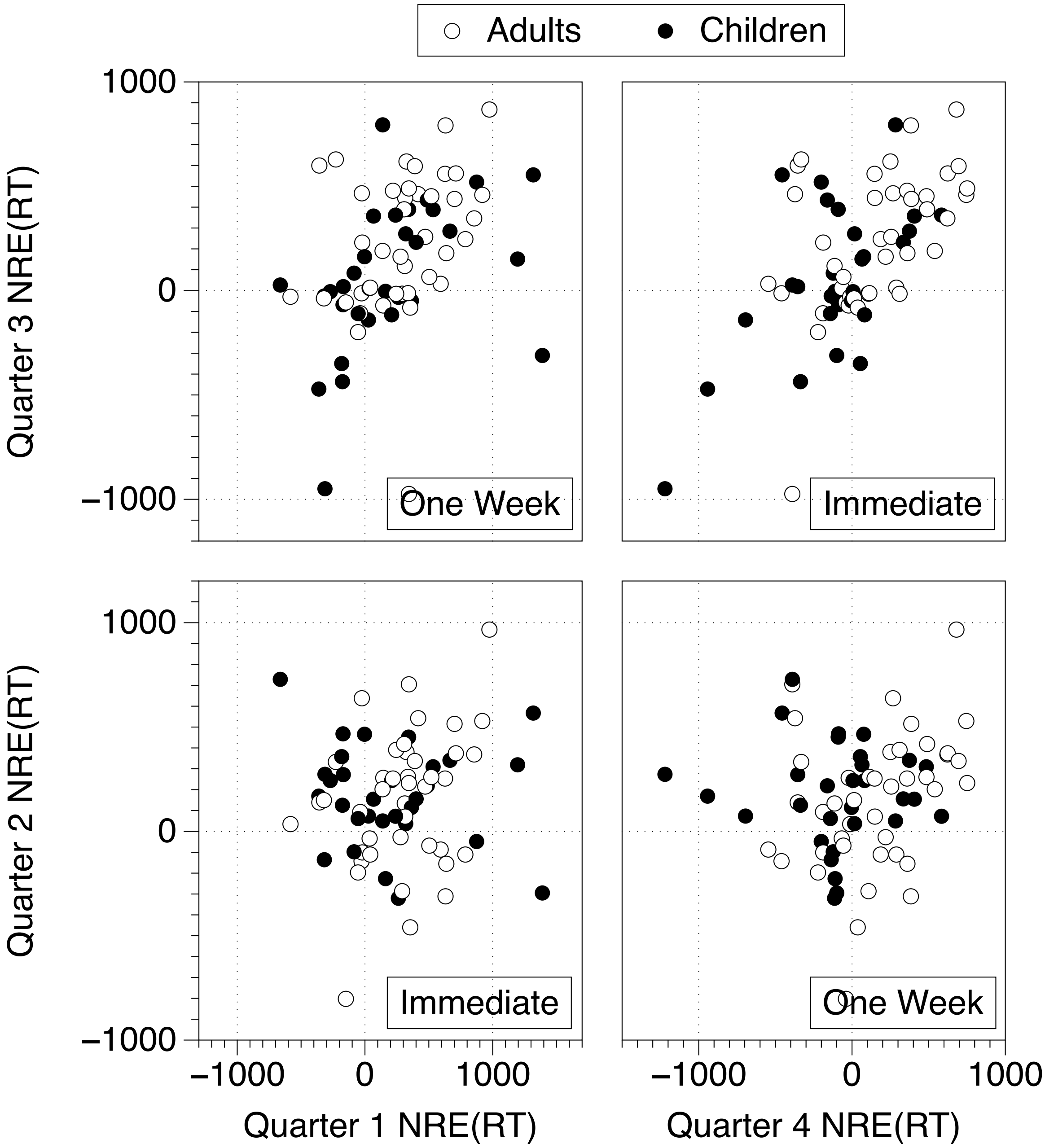


Figure 7. The test-retest reliability of the NRE(RT) index.


○ Adults
● Children
Quarter 3 NRE(RT)
Quarter 2 NRE(RT)
One Week
Immediate
Immediate
One Week
Quarter 1 NRE(RT)
Quarter 4 NRE(RT)