alternate-form reliability – The degree to which two or more versions of the same test correlate with one another. In clinical studies in which a given function is going to be tested more than once over time, it may be important to have more than one version of a given instrument and to show that these different versions are reasonably equivalent. Alternate form reliability is particularly important in testing cognitive functions such as memory. For example, if memory is tested through the learning of a list or words, follow-up testing should use a different list of words of equivalent difficulty. If the same list of words were used on the second testing, scores would be inflated due to the subject's prior knowledge of the words. These practice effects can be reduced but not completely eliminated by using alternate forms since mere familiarity with a task can boost scores even if the actual content changes.

auditory information processing speed – The ability to attend to orally presented information, process the information in memory, and then formulate a response in a timely fashion.

attention – The ability to focus on a specific task for a sustained period of time while remaining reasonably free of distraction.

battery – A group of individual assessment instruments administered as part of a comprehensive evaluation. Although the different instruments may be interrelated, each instrument can generally be administered, scored, and interpreted independently. Some assessment batteries allow for the generation of one or more composite scores that combine information from more than one of the instruments making up the battery.

bimodal distribution – When large numbers of patients are tested with a given instrument, the scores are often plotted on a graph in which the horizontal axis represents the scores (from low to high) and the vertical axis represents the number of patients obtaining a given score. This graph may take the form of a "bell-shaped" curve in which the distribution of scores forms a single, rounded peak in the middle of the graph and then gently tapers toward either extreme of the scale, i.e., like a bell. This shape comes about because the scores in the middle of the scale are the most common whereas scores at either extreme are much less common. A bimodal distribution in contrast has two peaks with a pronounced dip in-between. A bimodal distribution may result from an undesirable characteristic of a scale in which different scores along the scale do not represent equal intervals. In contrast, an equal interval scale is more likely to have a bell-shaped curve.

composite score (or scale) – A score based on combining scores from two or more individual items or scales. In many cases a composite score is created by simply adding individual item scores or scale scores together. For example, a 3-

item composite with a score of 9 could be created by adding together individual item scores of 3, 1, and 5. However, there are many other more complex ways to create composite scores. When a composite score is created, it is important to evaluate the internal consistency of the composite.

concurrent validity – A type of criterion-related validity in which the independent criterion is measured at the same time as the instrument under study.

construct – A characteristic of a person that is the object of measurement. The term construct generally refers to some phenomenon that is abstract and theoretical and that cannot be directly observed such as "depression." In order to measure a theoretical construct, one utilizes one or more indicators that are thought to be related to the construct. These indicators may take the form of one or more individual items on a multi-item scale. Many theoretical constructs are multifaceted and therefore consist of more than one domain. See also construct validity.

construct validity – The extent to which an instrument adequately measures a theoretical construct. Construct validity does not refer to a particular type of statistical test but rather to the accumulation of a variety of different types of evidence that may include content validity, criterion-related validity, convergent validity, discriminant validity and others. In order to establish construct-validity the investigator, based on theory, first delineates the meaning of a construct and its predicted relationship to other constructs some of which are similar and some different from the construct of interest. These predictions are then tested empirically and the results interpreted. The process of establishing convergent and discrimiant validity is sometimes used synonymously for construct validity however, construct validity encompasses a broader set of methods.

content validity – The extent to which an instrument includes a representative sample of the content of a construct. For example, an instrument designed to measure neurological impairment that only includes items concerned with brainstem function would have poor content validity.

convergent validity – The extent to which a measure is correlated with one or more other measures of the same or similar constructs. One common way to assess convergent validity is through criterion-related validity.

correlation – The extent to which two or more variables are associated with one another, generally used in the context of two variables. Correlation can be positive (as one variable increases, the other also increases) or negative (as one variable increases, the other decreases). There are a wide variety of methods for measuring correlation including intraclass correlation coefficients, the Pearson product-moment correlation coefficient, and the Spearman rank-order correlation.

criterion-related validity – The degree to which one measure is correlated with another, independent measure of the same construct. An example of criterion-related validity might be showing that a self-report questionnaire correlates highly with a diagnosis of depression established by a psychiatrist.

Cronbach's alpha – A commonly used statistical technique for evaluating the internal consistency of a scale in which several numerical items are added together to arrive at a composite score. It may be interpreted as the average correlation between the two halves of a multi-item instrument if that instrument were to be split into all possible combinations of two halves. A high alpha indicates good internal consistency and suggests that there is at least one fairly homogeneous dimension underlying the scores on the instrument.

domains – A theoretical construct such as depression may be multi-faceted and thus consist of several domains all of which are related in some way to the construct. For example the theoretical construct of depression consists of several domains including dysphoria, anhedonia, etc.

divergent (discriminant) validity – The demonstration that a measure is weakly or not at all related to a construct that theory predicts should indeed be different. For example, a measure of verbal memory and a measure of bladder function should be relatively independent of one another.

face validity – The extent to which an assessment instrument subjectively appears to be measuring what it is supposed to measure. For example, an instrument that is supposed to measure fatigue might be said to have face validity if it contained questions concerning feeling tired. Although face validity helps to make a self-report questionnaire more acceptable to respondents it is not essential to ensure the overall validity of an instrument.

internal consistency – Internal consistency or internal consistency reliability is a concept applicable to multi-item composite scales. It refers to a desirable condition in a composite scale in which the individual items are mathematically associated with one another and thus all appear to be measuring the same construct. Achieving high internal consistency begins with the selection of items all of which are relevant to the particular construct to be measured. There are many statistical techniques used to evaluate internal consistency including Cronbach's alpha, item-total correlations, and split-half reliability.

inter-rater reliability – Sometimes referred to as inter-rater agreement, this is a type of reliability assessment in which a particular assessment is completed by two or more raters and these different ratings are then compared. A variety of different statistical techniques are used to evaluate inter-rater agreement

including simple correlation, intra-class correlation, weighted kappa, and others. Inter-rater reliability is an important issue for clinical rating scales in which clinical judgment is an important part of the assessment.

interval scale – A type of scale in which different scores represent an ordering, e.g., from highest to lowest, and in which all the steps along the scale represent the same interval, i.e., a difference between two points on one part of the scale is the same as a difference between two points in another part of the scale. An example of an interval scale is the Celsius temperature scale. On the Celsius scale, moving from 20 degrees to 30 degrees represents the same amount of difference in heat as moving from 50 degrees to 60 degrees. Strictly speaking this type of scale should be called an "equal interval scale." Compare to ordinal and nominal scales. Properly constructed composite scales may be regarded as interval scales even though the individual items making up the scale are ordinal in nature.

intraclass correlation coefficients – A group of related statistical techniques that are frequently used to assess inter-rater agreement. The intraclass coefficients are more appropriate than simple correlation coefficients for assessing inter-rater agreement. Unlike simple correlation coefficients, the intra-class methods are sensitive both to random error ("noise") and to systematic error (statistical bias). For example, if two neurologists score a group of patients and one of them always produces scores that are five points higher than the other neurologist (systematic error), simple correlation techniques would indicate that these two physicians are in agreement. However, the intraclass coefficients would accurately portray the extent of disagreement between them. Simple correlation techniques are only sensitive to random error ("noise"). Random error refers to "noise" in scores due to chance factors, e.g., a loud noise distracts a neurologist while she is evaluating a patient and thereby affects the score.

intra-rater reliability – Sometimes referred to as intra-rater agreement, this is a type of reliability assessment in which the same assessment is completed by the same rater on two or more occasions. These different ratings are then compared, generally by means of correlation. Since the same individual is completing both assessments, the rater's subsequent ratings are contaminated by knowledge of earlier ratings.

item-total correlation – A statistical technique that is often used to help evaluate the internal consistency of a composite scale. The score for a single item from a multi-item scale is correlated with the total score for the remaining items in the scale. Items found to have low correlations may be eliminated or may be combined with other items to form a different scale.

Kappa – A statistical technique used to evaluate the extent of agreement between two or more independent evaluations of a categorical variable. Kappa takes into account the extent of agreement that could be expected on the basis of chance. A common use for Kappa would be to evaluate the extent of agreement between two or more clinicians who independently generate diagnoses for the same set of patients, diagnosis being a typical example of a categorical variable. See also weighted kappa.

multidimensional – A multidimensional scale is one that measures more than one construct or more than one domain of a single construct. In many cases, the multidimensional nature of the scale will be expressed through the generation of several subscales.

neuropsychological testing – Standardized, objective tests designed to evaluate specific cognitive domains (e.g., memory and reasoning) and their implications for intellectual functioning in everyday life. Neuropsychological testing is generally administered as a battery of tests.

nominal scale – A nominal scale is one in which the different categories represent a classification of some type without the implication of ordering. Examples of nominal scales include diagnostic categories, ethnic groups, eye color, etc.

normative data – A dataset containing numerical information derived from some reference group. A patient or group of patients is evaluated using some assessment technique and the scores obtained are then compared to those obtained by the reference group. The reference group can be very broadly based such as “the U.S. population” or very narrowly defined such as “male neurologists over the age of 50.” Normative data is particularly useful in evaluating human abilities. For example, a patient’s score on a test of memory can be compared to normative data for the same test derived from subjects of the same age and educational level as that of the patient. It is then possible to determine the extent to which the patient’s score deviates from the average score for similar individuals.

ordinal scale – An ordinal scale is one in which the scores represent a rank ordering from lowest to highest. Unlike an interval scale, the steps in an ordinal scale are not equal. For example, restaurants are often rated using an ordinal scale consisting of stars. A 3-star restaurant is better than a 2-star restaurant. However, we cannot assume that going from 2 stars to three stars represents an equivalent difference in quality nor can we assume that all 2-star restaurants are of the same quality. Many ordinal scales are used in MS including the EDSS, the Ambulation Index, and the Disease Steps.

Pearson product-moment correlation coefficient – The most commonly used method for assessing the degree of correlation or association between two variables measured using an interval or ratio scale. The coefficient can range from –1 to +1. A value of zero indicates that the two variables are completely independent of one another. A value of –1 or +1 indicates a perfect association.

practice effects - The extent to which repeated testing of an individual with a given assessment method affects that individual's scores on that method, in general leading to better scores. Practice effects can be particularly significant in the measurement of cognitive abilities, especially if the same items are used on more than one occasion. However, practice effects can be an issue even if no test item is ever repeated. This is because to some extent, subjects will improve on a test with practice because they overcome the anxiety related to encountering a new task or because they master the strategy needed to do well on the task.

predictive-validity – A type of criterion-related validity in which the independent criterion is measured at a later time than the instrument under study.

prospective memory – The ability to recall tasks or events that are planned for the future. This may include repetitive behaviors such as remembering to take medications.

psychometrics – Psychometrics is the science of measurement as applied to psychological and social constructs such as intelligence, depression, etc. However, the methodologies associated with psychometrics, e.g., reliability and validity, can be applied more broadly to a wide array of measurement issues such as neurological function.

quality of life – Most broadly, quality of life refers to those aspects of life that are important to a person. Although there are individual differences in the extent to which people value particular aspects of life, within a given culture people appear to be more similar than different. For example in American society people place high value on personal safety, physical mobility, independence, etc. The concept of "health-related quality of life" is derived form the more general concept and refers to those aspects of life that are important to an individual and that may be affected, either in a positive or negative way, by health and illness.

ratio scale – A ratio scale is similar to an interval scale in that it is an equal interval scale. However, a ratio scale also has a true or non-arbitrary zero point. For example, the Celsius temperature scale is not a ratio scale since its zero point is arbitrary. In contrast, the Kelvin scale is a ratio scale because its zero point is a true zero. In general a ratio scale is the only type of scale for which one can discuss "percent change." For example, using the EDSS, it would be

meaningless to say that a change from a score of 6 to a score of 3 represents a 50% improvement. This is because the zero point on the EDSS is an arbitrary designation of normal neurological functioning. It would however be appropriate to say that someone whose weight dropped from 100 kilograms to 50 kilograms was then 50% lighter in weight.

reciprocal – The reciprocal of a number is another number such that the product of the two numbers will equal one. For example, the reciprocal of the number 3 is 1/3.

reliability – Reliability can be defined in a variety of ways. It is generally understood to be the extent to which a measure is stable or consistent and produces similar results when administered repeatedly. A more technical definition of reliability is that it is the proportion of "true" variation in scores derived from a particular measure. The total variation in any given score may be thought of as consisting of true variation (the variation of interest) and error variation (which includes random error as well as systematic error). True variation is that variation which actually reflects differences in the construct under study, e.g., the actual severity of neurological impairment. Random error refers to "noise" in the scores due to chance factors, e.g., a loud noise distracts a neurologist evaluating a patient's gait and thereby affects the score. Systematic error refers to bias that influences scores in a specific direction in a fairly consistent way, e.g., one neurologist in a group tends to rate all patients as being more disabled than do other neurologists in the group. There are many variations on the measurement of reliability including alternate-forms, internal consistency, inter-rater agreement, intra-rater agreement, and test-retest.

retrospective memory – The ability to recall information or events from the past. An example might be remembering the name of someone you have met recently.

sensitivity – In its most general use, the sensitivity of a scale refers to its ability to detect true variations or differences in the theoretical construct that is measured by the scale. Depending on the nature of the scale and the research question under study, an investigator may want a scale to be sensitive to changes over time in the same patients, to differences between patients at a single point in time, or to differences between patients and normal controls. In its more restricted use, sensitivity refers to the probability that a diagnostic technique will detect a particular disease or condition when it does indeed exist in a patient.

social support – A general term that refers to a variety of concepts related to the extent to which an individual experiences support and assistance from family, friends, and acquaintances. The concept of social support can encompass a variety of different experiences including assistance with the performance of

everyday activities, emotional support, affection, sharing of leisure activities, etc.

split-half reliability – A method of estimating reliability by correlating two halves of the same scale, often accomplished by dividing the test into odd vs. even items. This technique is limited by the fact that different ways of dividing the items can produce different correlation coefficients. This limitation is overcome by Cronbach's alpha which provides a single number showing the correlation between all possible half scales.

Spearman rank-order correlation – A correlation coefficient for ranked, i.e., ordinal, data in which the steps on the scale represent higher vs. lower values but the steps are not equal. Mathematically the Spearman rank-order correlation coefficient is actually the Pearson product-moment correlation coefficient applied to ranks.

subscales – Many measurement instruments are multi-dimensional and are designed to measure more than one construct or more than one domain of a single construct. In such instances subscales can be constructed in which the various items from a scale are grouped into subscales. Although a subscale could consist of a single item, in most cases subscales consist of multiple individual items that have been combined into a composite score.

test-retest reliability – A way of estimating the reliability of a scale in which individuals are administered the same scale on two different occasions and then the two scores are correlated. This method of evaluating reliability is appropriate only if the phenomenon that the scale measures is known to be stable over the interval between assessments. If the phenomenon being measured fluctuates substantially over time, then the test-retest paradigm may significantly underestimate reliability. In using test-retest reliability, the investigator needs to take into account the possibility of practice effects which can artificially inflate the estimate of reliability.

validity – The degree to which an assessment measures what it is supposed to measure.

weighted Kappa – A refinement of the kappa coefficient that takes into account the "seriousness" of disagreement among raters. For example if one rater classifies an MS patient as relapsing-remitting and another rater classifies the same patient as relapsing-progressive, this disagreement is not as "serious" as one in which the two ratings are relapsing-remitting vs. primary progressive.