Evaluation of techniques to identify beneficial effects of nutrition and natural products on cognitive function

Keith A Wesnes

This article considers the appropriate selection of cognitive tasks for research into the effects of nutrition and natural products on mental performance. It is recommended that tests appropriate to the domains of the cognitive function under investigation be used, and a set of criteria is provided to enable researchers to select appropriate tests and test systems for their research purposes. Research in this field is generally performed to establish whether products can produce beneficial effects on cognitive function, including optimizing development in pediatric populations, reversing existing cognitive deficits, preventing age-related declines, counteracting fatigue-based impairment, or simply improving performance above normal levels. The requirements of tests for these purposes are detailed, and evidence is presented that properly developed cognitive test procedures can be used in a range of normal and clinical populations, even in large, long-term studies and in a variety of cultures and languages.

© 2010 ILSI Europe

INTRODUCTION

Cognitive function relates to the mental abilities that enable individuals to perform activities of daily living. Many aspects of cognitive function are relatively stable and unaffected by, for example, aging, fatigue, drugs, or trauma; whereas other aspects, such as attention and memory, are variable by nature and highly susceptible to change. Tests of cognitive function assess how well various cognitive skills are operating in an individual at any particular time. Such evaluations require individuals to perform tasks involving one or more cognitive domains. Thus, if a researcher wished to assess memory, the test would involve memorization of information by a subject, and the outcome measure would reflect how well the subject retrieved the information. Likewise, to assess the ability to sustain attention, the test could involve monitoring a source of information to detect predefined target stimuli over a period of time, and the outcome measures would reflect the speed and accuracy of the detections.

Tens of thousands of cognitive tasks have been developed and used in clinical research over the last 150 years, but this has become more of a hindrance to progress than an advantage. Even a small change to the methodology of a test may alter the outcome parameters, and the application of diverse tests in a particular field makes it difficult to meaningfully relate the effects of one study to another. Most reviews of changes in cognitive function associated with particular drugs, treatments, or disease states begin with the disclaimer that the wide variety of tests and test variants employed in the field has made it difficult or sometimes even impossible to render definitive conclusions. Another unfortunate result is that the findings of such reviews are generally more qualitative in nature than quantitative.

In nutrition, as in many other fields, the large number of available tests can create a bewildering choice for researchers, and in this field there is no central authoritative and definitive guide for selecting the appropriate cognitive test(s) for any particular purpose. The problem can even become circular; take, for example, a...
situation in which a particular test has identified a cognitive failure in a population, and a treatment is initiated to treat the problem, using the test as the outcome measure. If the test is a well-validated instrument associated with a large database and established clinical relevance, this will not be a problem; however, if there is little consensus as to what the test assesses, the conclusions and general relevance of the research findings will be limited. It should also be noted that all tests can reflect one or more aspects of the current cognitive status of an individual, but the major requirement for nutrition research is to identify change in function, and many tests are neither designed for nor capable of assessing change with an acceptable degree of precision.

**DOMAINS OF COGNITIVE FUNCTION**

It is important that researchers in this field identify the appropriate domain of cognitive function to investigate. Although “cognition enhancement” is an acceptable generic term, as is “health promoting,” science and regulators require targets that are more specific and which respect the independence of different domains when considering specific claims. For example, in the medical field, why would one expect a drug that helps pulmonary function to also help the liver? This illustrates the limitation of global scores of cognition for nutritional claims and should serve as a guide for researchers seeking to assess specific target domains of function. There are a number of core cognitive domains that can be evaluated, including attention, information processing, reasoning, memory, motor control, problem solving, and executive function. Taking memory as an example, there are four major types: episodic or declarative memory, working memory, semantic memory, and procedural memory.1 As Budson and Price1 illustrate, few conditions are associated with impairments to semantic memory and procedural memory, whereas working and episodic memory are impaired in a wide variety of neurological, psychiatric, surgical, and medical conditions. This creates a rationale for directing testing toward working and episodic memory as potentially more fruitful areas to evaluate in novel conditions, and most test systems recognize this approach. Furthermore, tests specific to particular domains are ideal, when available, because they help facilitate the substantiation of any claims made on the basis of the research findings. The most specific tests are tests of attention tests, because well-designed tests of this type do not require aspects of memory or reasoning for task performance, and changes in performance can, thus, be relatively clearly attributed to effects on attentional processes. Because attention is important for the performance of any task, when seeking to evaluate other domains, it is useful to also assess attention so that the relative contribution to any effects of changes to attention can be established. For example, the Digit Span Test assesses working memory by measuring how many digits a subject can hold in memory at one time. Changes in attention could influence performance on this task, but this can only be established if attention tests are performed concurrently. Nonetheless, Digit Span is not a direct test of attention. Overall, most well-established test batteries include assessments of attention, working memory, episodic memory, motor control, and aspects of executive function.

**MINIMUM REQUIREMENTS FOR COGNITIVE TESTS**

To help researchers in this field identify appropriate assessments for their purposes, a set of criteria will be presented that any test or test system should fulfill before being considered “fit for the purpose” of the study or research program being undertaken. None of the requirements discussed in this section are “nice to haves”; rather, they are all minimum requirements and essential properties of tests. That is to say, it is the absence of established evidence to satisfy any single criterion that should encourage the potential user to look elsewhere. Some tests are validated extensively but, for example, notoriously insensitive to change or simply unreliable, and their continued use does not facilitate the advancement of knowledge in this field. Unless such decision-making in test selection becomes more widely applied, the proliferation of methods used in nutrition research will continue unabated, as it has in other areas.

**Validation**

Validation is central to test selection and, as with the other criteria to be discussed, is a necessary but by no means sufficient requirement in the selection process. This is one of the most misunderstood concepts in science, often confused with reliability, sensitivity, widespread use, utility, and other concepts but which relates to none of these. The fundamental basis of the validity of a test is simply that it measures what it purports to measure. Thus, for instance, if a test is designed to measure the ability to store and later recall a series of words, it is a valid test of word recall if it is demonstrated conclusively to measure this ability. For it to become a valid test of episodic memory (the overall cognitive domain of which word recall is a part), the information generated by the test would need to show the same general patterns as other types of information from tests that have already been established as episodic verbal memory tests. Although the example of a word recall task is relatively easy to grasp, consider for example the construct validity
of using a test such as the Trail Making task as a measure of executive function. This is far less easy to establish, because of the variety of definitions of executive function and the numerous aspects of cognitive function involved in the performance of the Trail Making task. The probable answer is that Trail Making measures some aspects of executive function, but it also measures several aspects of cognitive function (e.g., attention, working memory, and motor control). Therefore, changes in performance on this test may not definitively be interpreted as changes to executive function, because other aspects of cognitive function may also have been involved. This aspect of validity, often termed construct validity, is widely recognized as the ultimate demonstration of the validity of a test (or test system) and generally involves complex statistical procedures, including cluster and factor analysis. Factor analysis can also be applied to the various measures of a test battery to establish the factor structure, and if tests believed to measure different aspects of function (e.g., attention and working and episodic memory) can be shown to load on independent factors (e.g., to cluster), this is an important step in establishing the construct validity of the test system.²,³

Other aspects of test validation that also need to be addressed are face validity, criterion validity, and predictive validity. Face validity refers to the purpose of a test being apparent by its nature, as with the example of word recall above. This is not an essential property, but it can facilitate the adherence of volunteers or patients when asked to perform a task because they can easily grasp why they are required to perform it. Criterion validity (sometimes termed concurrent validity) involves the demonstration that the scores from the test correlate with other established tests of the same aspect of cognition, providing evidence that the outcome measures move in a direction similar to those of the other tests. Correlation may not necessarily imply that the relationship is more direct, but correlations that account for a satisfactory amount of the variance yield some evidence that tests may be assessing the same aspects of cognitive function. Predictive validity is also a useful property of a test, for example, when used to identify an early sign of cognitive deterioration, which will proceed to result in clinical pathology such as Alzheimer’s disease (AD), or when used to determine whether an effect identified in volunteers will translate to an effect in a patient population.

Sensitivity

The governing principle here is simple to understand and easy to determine; when attempting to identify an effect, either positive or negative, researchers should only adopt a test that has previously been shown to detect such a change in cognitive function reliably, and ideally on more than one occasion. If the test is new, it should be used alongside other established tests in the field to assess its future utility. Many tests have the ability to discriminate between different groups or populations; while this “cross-sectional” ability to discriminate is a useful feature, this alone does not mean that the test will be suitable for repeated administration and be sensitive to differences that occur between groups or treatments over time. Thus, when seeking tests to determine change over time, researchers should be cautious about any test that has been shown to have only such cross-sectional sensitivity.

Reliability of tests over repeated administration.

Again, this is a construct that is simple to understand and easy to determine; a test is reliable if the scores it yields remain stable when administered repeatedly to individuals over occasions when there is no reason to expect a change in the particular ability that the test is designed to assess. Correlational measures are widely used, but they are inadequate alone, because high correlations between repeated administrations will often occur in tests that show large practice or training effects. The ideal demonstration of test–retest reliability involves stability of the scores on repeated administration, together with a reasonable degree of correlation between the repetitions. It is generally the case that tests involving complex strategies to perform, and/or a range of aspects of cognitive function, and/or skills that are subject to procedural learning often show training effects. Such training effects may be further increased if certain tests do not have parallel forms and the participant therefore performs better when retested simply because of having, for example, remembered the stimuli used in memory testing. The absence of parallel forms of certain tests greatly limits their usefulness in nutrition studies (or in any field) in which the principal object is to identify change.

It has long been recognized that such practice (training) effects exist with many cognitive and other tests⁴ and that these can compromise the ability to reliably identify effects of change in clinical trials.⁵ A number of factors contribute to this phenomenon in addition to simple “learning effects”; these include procedural learning (e.g., the unconscious improvement in test performance that occurs with repetition), the individual’s full understanding of the test requirements, initial “test-anxiety” that fades as familiarity with the test requirements increases, and the development of strategies required to perform tasks of greater complexity. Prestudy training of volunteers and patients is essential in nutrition (and all) studies in order to reduce these effects. In volunteers, it has been found that the training effects for many less complex tasks tend to plateau after four repetitions. Some tests (e.g., versions of choice reaction time tests) do not show
notable practice effects, although it is still good practice to train volunteers in order to minimize variability. This has led to the recommendation that four training sessions for each test should be conducted before the first day of testing in clinical trials. In trials with patients, it may not always be feasible to conduct four prestudy training test sessions, but at least two sessions is generally a useful minimum requirement.

Another important consideration for nutrition research is whether practice effects exist in long-term trials, because the training effects may dissipate if testing is separated by long periods, such as 6 months or a year. Wilson et al. conducted a 6-year study in an elderly population in which a range of traditional neuropsychological tests were administered annually. Surprisingly, practice effects persisted until at least the third year on many measures, making a strong case for prestudy training and the use in such studies of tests that do not show such training effects. In one long-term study, 257 hypertensive patients aged 70 years and older performed a range of computerized tests of attention and working and episodic memory yearly over a 5-year period. The participants were trained on the tests before entry into the study, and no practice effects were seen over the 5 years of the study. The participants receiving a placebo showed declines in global measures of attention and episodic memory over the study period, while the participants who were administered the study treatment (the antihypertensive candesartan) showed significantly less decline. Such work is important for the planning and conduct of nutrition trials because it demonstrates that long-term trials of nutrition are feasible if the appropriate types of cognitive tests are used. This trial was a substudy of the Study on Cognition and Prognosis in the Elderly (SCOPE) – an international multicenter trial conducted in 4,937 hypertensive patients aged 70–89 years. The primary cognitive outcome of the SCOPE trial was the Mini-Mental State Examination (MMSE) score; although no beneficial effects of the study compound (candesartan) were seen on the MMSE score in this large population, in the sub-study of 257 patients, computerized testing revealed significant benefits of the compound in several key domains, thus illustrating the sensitivity such testing can bring to long-term trials.

Utility

Experienced and properly qualified psychologists are rarely widely available in clinical trials, and when they are not, tests should ideally be simple to administer and easy for participants to understand. Furthermore, tests need to be demanding if they are to properly evaluate the aspect of function under scrutiny, but they should not be onerous (e.g., excessively long or threatening, as when they involve negative feedback) because this will reduce compliance. Moreover, tests that induce anxiety may give misleading results, because a substance that reduces anxiety may improve performance and thus be mistaken for a cognition enhancer. The instruments of choice for nutrition trials are tests or test systems with well-established and easy-to-administer instructions, both for test administration and scoring. Another important aspect of the utility of a test is its availability in various languages. If a researcher wishes to conduct trials in different countries, it is essential that the instructions be properly translated, and, if verbal material is used in the tasks, such as memory tests, these should be created in the appropriate language and not simply translated. Translation is unsatisfactory for word lists, because words have different frequencies of use in different countries and different numbers of syllables. When creating a set of parallel word lists for a memory test, it is important to balance the lists in terms of the frequency and number of syllables of the words, because both factors influence the ease with which words are learned. Frequency lists are available and should be used for this purpose (e.g., http://www.wordcount.org). Other aspects, such as the ability to visualize various words, are also important, and information is available for this. When such rules are applied, cross-language and cross-cultural stability can be satisfactorily achieved, even when Latin- (e.g., German) and character-based (e.g., simplified or traditional Chinese) languages are used in the same study.

OTHER IMPORTANT ATTRIBUTES OF TESTS AND TEST SYSTEMS

Normative databases

Established tests and test systems sometimes have normative databases containing data from the test’s target populations, and they are generally based on age and gender. These databases facilitate baseline testing in trials to establish whether or not the population is in the normal range; they also permit patient populations to be compared with equivalent groups of normal populations to enable the extent of cognitive deficits in the patient population to be identified. A further opportunity that is becoming more widely used in trials of cognition enhancers is the ability to identify the treatment response in terms of the degree to which the patients have moved toward the normal level for their age. Another application for such a database is its use in older volunteers to determine the degree to which the study compound may have ameliorated age-related declines.
Everyday relevance and clinical relevance

Although it is important to measure improvements in function precisely, it is also valuable to determine whether or not these are likely to have everyday benefits for the clinical population. Activities of daily living scales are used widely to measure the difficulties that people may experience conducting everyday tasks. The everyday relevance of tests and test systems can be determined using multiple regression techniques to relate the activities of daily living scores to performance on cognitive tests. These regression techniques control for extraneous factors and determine which of a variety of measures, including those from cognitive tests, can best predict the likelihood that patients will experience everyday problems. Computerized tests of attention have been shown to perform particularly well in this regard, and test systems that possess such relevance to activities of daily living scores have a clear advantage over tests that do not have this relevance established.

The clinical relevance of tests can be determined, for example, according to the degree to which test scores can be used to discriminate groups of people or to identify deficits in domains that are independently confirmed by clinicians. Furthermore, if groups of participants are stratified into disease severity stages, using for example the Global Deterioration Scale in dementia research, and cognitive tests can confirm such stratification, then this further establishes the clinical relevance of the tests.

Automation of cognitive tests

Automation of cognitive tests brings numerous advantages, and the most relevant to the area of cognition enhancement is improving the signal-to-noise ratio. The standardization that such testing can bring to test administration and the reduction of errors in scoring decreases noise (unwanted variability); however, the extra precision in assessment that millisecond resolution of response times can bring can also increase the signal. Furthermore, aspects of cognitive function can be assessed that cannot be measured using traditional pencil-and-paper measures. Major tests of attention such as simple and choice reaction time have always been automated, as have intensive vigilance tests such as the continuous performance test and digit vigilance tasks. When simple- and choice-reaction-time tasks are administered together, information can be gathered on important domains of attention and information processing in addition to the ability to focus attention that the speed of responding captures. In the last decade, cognitive processing times (also known as cognitive reaction times) and the variability of reaction times have been shown to provide crucial independent information that can differentiate, for example, types of dementia. Selective slowing of choice reaction time (e.g., slowed cognitive reaction time) has been shown to differentiate people with vascular dementia from those with AD, whereas greater reaction time variability in attention tasks differentiates people who have dementia with Lewy bodies (DLB) from those who have vascular dementia and AD with high specificity (86% and 98%, respectively). One study compared subjects with DLB, AD, Parkinson’s disease dementia (PDD), Parkinson’s disease (PD), and controls. Cognitive reaction time and reaction time variability were selectively disrupted in the PDD and DLB groups but not in the other groups. These deficits have now become recognized as hallmarks of PDD and DLB, and the consensus criteria for both types of dementia have included these deficits as part of the core clinical features of the conditions.

Furthermore, computerized tests of verbal and object recognition permit assessment of the time taken to retrieve the information from memory in addition to the accuracy of recognition. Traditional tests that cannot make this assessment have overlooked this important aspect of memory, but it is one that declines markedly and independently of accuracy with normal aging, and it is severely compromised in many debilitating diseases such as dementia. Such slowed speed of information retrieval is an early characteristic of mild cognitive impairment, which can respond to pharmacological treatment. Automation provides the same benefits for tests of ability to retain information in working memory, because the role of working memory is to facilitate the performance of ongoing tasks. Clearly, it is not just the ability to retrieve the information correctly that is important, but also the time this task requires; this is something traditional tests such Digit Span cannot assess and is something that, again, is impaired with dementia. A further important benefit of assessing speed is that it permits trade-offs between speed and accuracy to be identified, which helps avoid the misinterpretation of study findings.

FUTURE DIRECTIONS FOR COGNITIVE TESTING IN NUTRITION RESEARCH

Although there are some similarities between the research aims in drug development and those in nutrition research, there are also important differences; the major one is that cognitive testing is widely employed in drug development to screen for unwanted cognitive impairment or interactions with other drugs or alcohol. Such impairments can be detected even with relatively insensitive pencil-and-paper tests, and this may be the reason such testing persists in many trials. However, in nutrition research, the object is most often to detect benefits to function, by improving healthy individuals or by prevent-
ing and even reversing the declines seen with aging or disease. Computerized tests have proven useful for such work in drug development, particularly test systems designed specifically to satisfy the various criteria described in this article and which measure a range of aspects of cognitive function. Such test systems have helped researchers identify benefits in a range of conditions, for example, attention deficit hyperactivity disorder, AD, hypertension, and epilepsy. Furthermore, in nutrition studies, such computerized testing has shown sensitivity to the effects of breakfast, energy drinks, and a range of natural substances. The benefits of particular measures can be detected in such trials within hours, e.g., for breakfast and energy drinks, or over months or even years.

CONCLUSION

Researchers in the field of nutrition who wish to identify changes in cognitive function, particularly enhancement, face a perplexing choice in the selection of the appropriate cognitive tests or test batteries for their studies. This article describes a number of essential properties of suitable test procedures based on the experience of the author over the last 4 decades. It has been argued that utility, reliability, sensitivity, and validity are the independent minimum requirements that need to be satisfied before a test or test system can be considered to be “fit for purpose” for the aims of detecting change in cognitive function. Clinical relevance, everyday behavioral relevance, and normative databases are also highly desirable properties of tests and test systems. Automated tests have significant advantages over traditional pencil-and-paper and stopwatch testing procedures in terms of helping to reduce unwanted noise and independently enhancing signal strength, both of which improve the signal-to-noise ratio in clinical trials. Of relevance is a recent study in which the test administrators and the subjects (community-dwelling individuals aged 85 years and older) were asked to rate computerized tests (e.g., tests of simple and choice reaction time, digit vigilance, and word recognition) and nonautomated tests (e.g., the Wechsler Digit-Symbol test and word list learning tasks) for acceptability. The administrators and the participants rated the computerized tests as more acceptable, and only 91% of the sample could complete the pencil-and-paper tasks, whereas 100% were able to complete the computerized tests.

In conclusion, evidence suggests that appropriately developed automated test systems can overcome the majority of the widely perceived potential barriers to the use of such procedures in nutritional trials, even in large-scale, long-term, multicenter international trials, providing data of the quality required by regulatory agencies in the pharmaceutical arena, as well as criteria set by equivalent bodies in the nutrition field (e.g., Process for the Assessment of Scientific Support for Claims on Foods). The preservation, restoration, and optimization of cognitive function are widely sought in most cultures, and nutrition can undoubtedly play an important role in helping to achieve these aims. In order to determine which nutritional strategies are most effective, it is necessary that the trials performed be conducted using appropriate standards and appropriate test instruments.

Acknowledgments

Declaration of interest. Prof. Wesnes is an employee of a company that provides cognitive testing services to the pharmaceutical and nutritional industry. He received a small honorarium for writing this article.

This work was commissioned by the Nutrition and Mental Performance Task Force of the European branch of the International Life Sciences Institute (ILSI Europe). Industry members of this task force are Abbott Nutrition, Barilla G. & R. Fratelli, Coca-Cola Europe, Danone, Dr Willmar Schwabe, DSM, Friesland Campina, Kellogg Europe, Kraft Foods, Martek Biosciences Corporation, Naturex, Nestlé, PepsiCo International, Pfizer, Roquette, Soremartec – Ferrero Group, Südzucker/BENEO Group, Unilever. For further information about ILSI Europe, please call +32-2-771-00-14 or email: info@ilsieurope.be. The opinions expressed herein are those of the authors and do not necessarily represent the views of ILSI Europe. The coordinator for this supplement was Ms Agnes Meheust, ILSI Europe.

REFERENCES