Cognitive and Psychomotor Performance Tests and Experiment Design in Multiple Chemical Sensitivity

Anthony Wetherell

Defence Evaluation and Research Agency, Protection and Life Sciences Division, Chemical and Biological Defence Human Studies Group, Porton Down, Salisbury, United Kingdom

People suffering from multiple chemical sensitivity (MCS) complain of a variety of symptoms that could impair cognitive and psychomotor function either directly or indirectly. This paper discusses the use of cognitive and psychomotor performance tests together with some experiment designs that could be considered for use to assess fitness of MCS sufferers for work or the efficacy of diagnostic, preventative, or therapeutic measures. The tests could also contribute to the body of objective information on MCS and help sway the opinion of those who are dubious of its authenticity. The credentials of cognitive and psychomotor performance tests are derived from their successful use in studying the effects of drugs, and the types of tests are illustrated by describing those used by the United Kingdom Defence Evaluation and Research Agency Chemical and Biological Defence Human Studies Group, which has been involved in the assessment of drugs and chemicals on work performance for many years. The tests include mathematical, verbal and spatial processing, tracking, reaction time, attention and vigilance, and memory tests. The discussion of experiment designs includes both repeated measures and parallel groups designs together with their advantages and disadvantages and some suggested modifications to accommodate the particular problems posed by MCS. — Environ Health Perspect 105(Suppl 2):495–503 (1997)

Key words: multiple chemical sensitivity, cognitive function, psychomotor function, performance tests, performance models, psychometrics, sensitivity, reliability, validity, experiment design, parallel groups designs, repeated measures designs, Latin square designs, drug effects, psychopharmacology

Introduction

Multiple chemical sensitivity (MCS) describes an intolerance or a hypersensitivity to chemicals or other agents in the environment. The condition can arise from an acute or chronic exposure to one chemical or agent; it is marked by heightened sensitivity to subsequent, very low levels of exposure and often expands to include other chemicals (hence multiple sensitivity).

The range of possible chemicals that give rise to MCS is vast, but common ones include fuel and oil fumes and combustion products, perfumes or colognes, cleaning agents, building and decorating materials, and foodstuffs or additives.

Some MCS sufferers know which chemicals are responsible for their condition, but many do not—they simply feel ill and might or might not suspect the cause of their discomfort. Symptoms vary considerably among individuals, and often are difficult to define, which leads some authorities to doubt the authenticity of MCS. Several organ or tissue systems can be involved, usually the nervous, respiratory, gastrointestinal, and musculoskeletal systems. Symptoms typically include headache, fatigue, joint and muscle problems, irritation to the eyes, ears, nose, throat or skin, general malaise, difficulty concentrating, and poor memory. These symptoms vary in severity and can be disabling. Whether or not MCS is a recognized medical condition, there is no doubt that sufferers feel ill, often are aggrieved when their complaints are dismissed, and often feel unable to work.

Studies of MCS have mostly been concerned with elucidating the mechanisms by which the chemicals have their effects, identifying and treating MCS sufferers, or assessing the veracity of their claims. The first type of study commonly uses techniques such as neuropsychology, immunology, psychoneuromimmunology, neurophysiology, and nasal pathology and olfaction. The second and third types of study commonly involve removing the suspect chemical or chemicals from the patient or removing patients from the chemicals, usually by confining them in a specially built, chemically clean environment. In either case, it is important that the studies be properly designed and controlled and that the outcome measures of the effects of the procedure are sensitive, reliable, and valid. Bad study design and measurements are unlikely to find answers, and are just as unlikely to sway the opinion of those who doubt the authenticity of MCS.

Cognitive and psychomotor performance tests, as distinct from neuropsychological tests, have been little used so far in evaluating MCS but have been widely used to assess the effects of environmental stressors, mostly drugs. These types of test have three advantages for MCS patients. First, they are based on accepted models of real-life skills, and thus can be used as objective assessments of the patients’ abilities to work. Second, they are very sensitive and can easily detect and measure effects of exposure or the effectiveness of diagnostic, preventative, or therapeutic measures. Third, the tests are objective, and their use, together with carefully controlled study designs, could help to determine whether MCS is a genuine illness.

Cognitive and Psychomotor Performance Tests

Cognitive function refers to an individual’s ability to think and reason in terms of temporal and spatial relationships and in symbols such as words and numbers. Psychomotor function (sometimes called perceptual-motor function) refers to an individual’s ability to coordinate timely and appropriate responses to stimuli. Cognitive and psychomotor functions are not discrete; they overlap to the extent that the stimuli required thought. For example, a simple reaction time test is
usually considered a psychomotor test, but if the stimuli are complex and require decisions about how to respond, then the test becomes more cognitive.

A distinction should be drawn between cognitive and psychomotor performance tests on the one hand, and neuropsychological and other types of psychological tests on the other hand. Both are intended to elucidate cognitive and psychomotor function and, thus, there is a large amount of overlap, e.g., reaction time is used as both a performance test and a neuropsychological test. The difference is in the basis and use of the tests. Neuropsychological tests are based mostly on neurological principles; they are almost always purchased from suppliers, administered in standardized forms, and are mostly diagnostic in purpose. Because neuropsychological tests are standardized, the score of any individual on any occasion can be compared with norms and the degree of abnormality determined. Neuropsychological tests have been used frequently in MCS studies but generally have not shown any consistent impairment in MCS sufferers.

In contrast, cognitive and psychomotor performance tests are based mostly on performance models such as factor analysis, resource, and information processing models. Details of these models can be found in any experimental psychology text, and their relative merits have been discussed earlier (1). Performance tests are not normally purchased. They are derived mostly from the experimental psychology literature and are in the public domain; thus, anybody can use them. Performance tests are not diagnostic but predict real-life performance to the extent that they have construct and criterion validity. Most performance tests have a high degree of construct validity, but it must be admitted that many have questionable criterion validity.

Performance tests generally are not standardized; they are almost always adapted to improve their sensitivity for the purposes for which they are used. Some researchers say that standardization is desirable so results can be compared among laboratories; others say that standardization reduces the tests' sensitivity. Both are right, but the approaches are mutually exclusive. Some attempts at standardization have been made, e.g., the NATO Advisory Group on Aerospace Research and Development (AGARD) Standardised Tests for Research into Environmental Stressors (STRES) Battery (2,3), but the tradition of individual laboratories developing tests for their own purposes is too well established to be easily abandoned.

Because performance tests are not standardized, they generally have no norms. This has advantages and disadvantages. A disadvantage is that the tests cannot be used to determine whether a score is abnormal. However, determining whether a score is normal or abnormal depends on the sensitivity and reliability of the test in discriminating between normal and abnormal individuals, and on the reliability and validity of the norms. Norms should not be used uncritically but considered in terms of their size or coverage, e.g., age, gender, socioeconomic grouping, occupation, culture, or nationality.

There actually is a small number of cognitive and psychomotor abilities, according to the classifications and taxonomies proposed (4-8), but people use these abilities in a variety of ways and there is a staggering number and variety of tests. Most tests have been used in studies of drug effects (9-11). To date, there have been only a few studies of the effects of industrial and environmental chemicals using cognitive and psychomotor performance tests, as distinct from neuropsychological tests (12,13), but there is no reason to suppose that tests for drug effects would not also be suitable for testing for chemical effects.

All psychological tests should satisfy three basic psychometric criteria: sensitivity, reliability, and validity (1). Sensitivity concerns whether the test can measure anything at all; a test may be sensitive to one stress but insensitive to another. Reliability concerns whether the test makes its measurements consistently; an unreliable test will return different scores each time it is used, a situation that can be mistaken for effects of, for example, a chemical. Validity concerns whether the test measures what it claims to measure; an invalid test may be sensitive and reliable but measures the wrong thing.

These principles are applied in many areas of psychology such as personality, intelligence, and clinical and occupational testing but are not applied as often as they should be to cognitive and psychomotor performance assessment. This is particularly true in psychopharmacology, where the literature abounds with elegant, imaginative, and ingenious tests of doubtful use. Validity also is related to the purpose of the test: a test may be valid for one purpose but not for another. Cognitive and psychomotor performance tests are valid mostly for the study of drug effects, but their validity with respect to MCS cannot be determined until they are tried.

It has been mentioned that performance tests are based on performance models. These models have been discussed elsewhere (1), but a brief mention is appropriate here. There are many performance models, but they can be classified into three basic types: factor analysis, resource, and information processing. Factor analysis models attempt to identify basic factors underlying performance by correlating performance on various tests and performance in real life. The resource model postulates that humans have one or more limited pools of resources that can be allocated to the performance of tasks. Information processing is a fashionable but widely abused term sometimes applied to any procedure in which information is processed; the term is more correctly applied to procedures that attempt to elucidate the mechanisms by which information is processed between stimulus and response. Perhaps the most useful variant of information processing is the stage processing model, which holds that humans process information through several serial stages such as detection, recognition, decision, and response. These stages can be isolated using additive factor methods (1) and the loci of effects, e.g., of drugs, determined.

Tests based on factor analysis and resource models are phenomenological in that they represent and often resemble skills and behavioral phenomena that humans exhibit in real life. On the other hand, tests based on the stage processing model do not necessarily resemble real life; rather, they attempt to isolate the various stages involved and to assess their contribution to the overall performance of the task.

The types of test in common use are illustrated by describing those used at the United Kingdom Defence Evaluation and Research Agency Chemical and Biological Defence (CBD) Human Studies Group. The task of the Human Studies Group is to assess the effects of drugs, chemicals, and other environmental stressors on humans—until recently just for the military but now for civilian industry as well. Since real-life jobs are too numerous and complex, or too difficult, to study individually, CBD has used most of the cognitive and psychomotor tests available and continues to develop new ones to keep pace with increasing technology and work practices.

The tests listed below exist in a great many versions; the ones described are those in current use. CBD has found them to be valid, reliable, to have a good track record.
in assessing environmental stressors, and to be the best for assessing the effects on work performance of a variety of drugs including anticholinergics such as atropine (14,15) and hyoscine (16,17), benzodiazepines such as diazepam (14), anticholinesterases such as sarin (18), pyridostigmine (19) and physostigmine (17,20), anxiolytics such as ondansetron and granisetron (21), and antibiotics such as doxycycline and ciprofloxacin (A Wetherell, unpublished data). The tests have also proven useful for studying other stressors such as fatigue, sleep deprivation, and protective clothing (22).

CBD’s tests are all in the public domain and come from a variety of sources such as the cognitive, clinical and experimental psychology literatures, the batteries developed by the U.S. armed services, the NATO AGARD STRES Battery (2,3), and some tests developed at CBD. Not all the tests are used all the time; they are selected according to the particular drug, or other stressor, and work situation at issue. All the tests can be administered repeatedly, to monitor the time-course of effect; different versions are produced by random or pseudorandom generation of stimuli.

Unless stated otherwise, all tests are presented and scored by computer and last 3 min; scores are the number of problems attempted, the number of correct answers given, and response times, or derivations of these. References refer to the originator(s) of the test or to examples of the first or early use of the test. Where no reference is given, the test has been in use for so long that its origin is not known.

Mathematical Processing

**Numerical Processing (2,3,23).** A series of problems is presented, each problem consisting of three digits and two operators (+ or −), e.g., 6 + 4 – 3. Subjects say whether the answer is greater or less than 5. Problems are designed so the answer never actually equals 5.

**Number Facility (24).** A series of problems is presented, each problem consisting of three one- or two-digit numbers arranged vertically with a box at the bottom. Subjects have to sum the numbers and insert the answer in the box.

Logical Reasoning

**Original Version (25).** A series of sentences, each followed by a pair of letters, e.g., AB or BA, is presented. The sentence describes the order of the letters, e.g., A follows B, B is preceded by A, A is not followed by B, and subjects have to say whether the statement is true or false. For example, B does not follow A–B is false. **AGARD STRES Version (2,3).** A series of pairs of sentences, each followed by three symbols, #&*, is presented. The sentences describe the order of the symbols, e.g., & before #, after * & # & * . If both sentences are true (or both false) with respect to the three symbols, subjects press a key signifying same, otherwise, they press a key signifying different. In the above example, both sentences are false. AGARD adopted this version because the simpler syntax is more language-fair, and the symbols have no inherent order. Two sentences with three symbols are used because using the simpler syntax of one sentence with two symbols would have been too easy.

**Spatial Processing**

**Manikin (26,27).** A front or back view of a human, rotated at any angle, and holding a flag in one hand is presented. Subjects have to specify which hand is holding the flag. **Histograms (2,3,23,28,29).** A four-bar histogram is presented for 3 sec, followed by a blank screen for 1 sec, followed by a second histogram rotated by 90 or 270°. The subject must say whether the two histograms are the same or different.

**Tracking**

**Pursuit Tracking.** The subject, using a joystick, tries to keep a cursor on a moving target. The test lasts for 3 min, and the measurement is the root mean square error.

**Unstable Tracking (2,3,23,30).** The subject uses a joystick or a mouse to keep a horizontally moving cursor on a fixed target. The test is set up so the cursor accelerates away from the target, requiring subjects to increase their control movements with increasing distance. This is analogous to balancing a pool cue vertically on the end of a finger.

**Reaction Time**

**Simple Reaction Time.** Subjects press a key as quickly as possible after a stimulus, usually a black square or disk on the screen, or a beep from the speakers. Reaction time is usually taken as the mean of several attempts, to minimize errors due, for example, to distraction.

**Choice Reaction Time.** Subjects press one of several keys as quickly as possible after various stimuli. Reaction time is usually taken as the mean of several attempts, to minimize errors due, for example, to distraction.

**Complex Reaction Time (2,3,23,31).** This test is based on the stage processing model and is designed to identify the locus of a drug effect. A tautologous digit (2, 3, 4, or 5) is presented on the left or right of the computer screen and subjects press one of four keys as quickly as possible. The test has six parts, each lasting 3 min and each designed to cover a particular stage of information processing—encoding, motor programming, motor activation, response selection, and response execution. This is achieved by varying the stimulus quality (easy or difficult to recognize), response complexity (single or triple key presses), time uncertainty (regular vs unpredictable interstimulus intervals), and stimulus–response compatibility (response key on same side as or different side from the stimulus).

**Attention/Vigilance**

Attention is a driver for other functions, but in test terms it is usually thought of as the ability to detect fairly frequent targets in a matrix of rapidly presented, repetitive stimuli. In contrast, vigilance is the detection of uncertain or infrequent stimuli over a prolonged period. Many so-called attention/vigilance tests involve higher cognitive functions than simply detection, so the term is often used to include tests that are difficult to describe in any other way.

**Letter Cancellation.** Matrices of random letters are presented, and subjects cross out or mark certain letters.

**Serial Response (32).** A row of five outlined squares is presented, corresponding to the keys 1 to 5 on the keyboard. Subjects "chase" a black square that appears at random in one of the outlined squares by pressing the appropriate key; each key press causes the black square to disappear and reappear.

**Focused Attention (33).** Three warning crosses are presented, one in the middle of the screen, the other two either close to it or close to the edges of the screen. The middle cross is replaced by a target letter (A or B) and the other crosses by asterisks, the same letter as the target, or the other letter. Subjects respond to the target letter by pressing an appropriate key.

**Search (33).** Two warning crosses are presented close to the middle or close to the edges of the screen. One cross is then replaced by a target letter (A or B) and the other is either replaced by a digit or disappears. Subjects respond to the target letter by pressing an appropriate key.

**Display Monitoring (23).** This test was designed to assess performance in
process control situations. Subjects watch the display of a scale and a moving pointer. At random intervals, the pointer tends to stay in one-half its scale. Subjects must report when this occurs. The load can be varied by having subjects watch one, two, or four displays at the same time.

**Vigilance.** Several auditory and visual vigilance tests are used, all requiring subjects to detect signals, or targets, in noise. Typically, the noise can be white noise, tones of various lengths and frequencies, or strings of digits, letters, or other symbols. The size of the population of symbols will affect subjects' performances, e.g., letters are more difficult to detect than digits because there are 26 possible letters and only 10 digits. The targets can be tones of different lengths or frequencies, or particular digits, letters or symbols, or groups of symbols. For example, one test consists of strings of digits, and subjects must press a key every time they see or hear three successive odd or even digits (34). The signal-to-noise ratio can be varied in terms of frequency of occurrence, intensity, and degree of similarity.

**Color–Word Naming (35).** Names of colors are presented, either in their own color or in different colors, e.g., the word green may be presented in green or in red, blue, yellow, etc. Subjects must name either the word or the color.

**Self-generation Tests**

These tests do not present stimuli but require subjects to self-generate responses. They are used mostly to measure perceptual–motor or cognitive load, and although they are sometimes used by themselves, they are often used as additional tasks in multitasking tests to measure reserve capacity or resource allocation.

**Interval Production (36).** Subjects must generate intervals, typically by tapping a finger, a foot, or by saying something at regular intervals, typically once a second. The actual regularity is measured: regularity is inversely proportional to perceptual–motor load.

**Random Generating (37).** Subjects must produce digits, letters, days of the week, or months of the year as randomly as possible. The degree of randomness is measured for single items and for groups of items as a function of the population of possible items: randomness is inversely proportional to mental load.

**Memory**

There are probably more memory tests than all other tests combined, since there are so many aspects of memory and because memory is involved in practically all cognitive and psychomotor functions. Indeed, some of the tests listed above involve so much memory that they might better be called memory tests. Memory tests generally present either meaningless information such as random digits or letters, or nonsense syllables or words, or meaningful information such as real words, sentences or short stories. At various times after presentation, subjects have to recall or recognize the information, either unaided or cued in various ways. Memory tests are affected particularly by learning, since that is what they are designed to measure. Most psychologists want to improve learning, but for a performance psychologist it is a considerable nuisance, since it confounds the results of repeated testing.

**Digit Span (38).** A set of digits, usually four, is presented one digit at a time. Immediately afterwards, subjects must recall the digits. If they succeed, a five-digit set is presented, and so on until they fail to recall the digits; then another attempt at a different set of the same length is allowed. If they succeed, they go on; if they fail, the test is ended and the score is the longest set of digits remembered. The test is often repeated, with subjects having to recall the digits in reverse order.

**Item Recall.** Lists of digits, letters, nonsense syllables, or real words are presented, and subjects must recall them. In one version (38) ten pairs of words are presented. In six of the pairs, the words are related to one another, e.g., eagle–bird; in the other four, the words are not related. Subjects are then presented with the first word of each pair and must recall the second.

**Memory Search (39–42).** This test is based on the stage processing model of performance, and can separate effects on memory from those on, for example, perception or response, which might be confused with memory effects. Sets of symbols, usually digits (target sets) are presented, each followed by a single probe digit. Subjects must say whether the probe digit is a member of the target set, and their accuracy and response times are measured. The target set size can be changed, the probe items can be made more difficult to see or recognize, and the response mechanisms can be changed to affect particular processing stages. **Shopping List.** A list of items is presented. Subjects are then given a box containing the items on the list together with the same number of other items and must pick out the items on the list. Handling real objects rather than remembering items in the abstract improves motivation and performance and is more representative of real-life situations.

**QRS Test (43).** This is a test of working memory based on the stage processing model. The letters Q, R, S, and T are presented randomly; subjects must count each occurrence of each letter and report the counts when asked. One, two, or all four letters may be presented, the starting value for each letter's count may be varied, and the incremental (or decremental) value varied to alter the load on the stages involved.

**Face Recognition.** Most memory tests use alphanumeric information and there is evidence that graphic or spatial information is processed by a separate system (17). A set of photographs of people or faces is presented and subjects must recognize them from a larger set.

**Incidental Memory.** Subjects are not given specific information to remember but are asked to recall incidental features of the test or the situation. One disadvantage with repeated use of this test is that subjects learn they will be asked to recall incidental features and begin to notice their surroundings more closely. This helps because in a controlled environment there is only a limited number of incidental features that can be used.

**Multitasking**

People in real life commonly have to attend to more than one task at a time. The above tests can be used in any combination, although information processing tests are most commonly combined with psychomotor tests such as tracking, attention/vigilance tests, or self-generation tests. The tests are chosen according to the combinations of tasks involved in particular jobs or to stress particular psychological functions that are susceptible to drug or other effects. Care must be taken to ensure that the sensory and motor modes do not conflict, or that they conflict in the way intended; people have a limited number of hands and fingers, which do not work completely independently.

**Experiment Design**

Cognitive and psychomotor performance tests are rarely standardized and rarely have any population norms. Thus, control, or baseline, data must be obtained for comparison each time the tests are used, and
this requires some form of study design. The procedures described are based on those commonly used in psychopharmacology; it is not intended that they all be used directly in MCS studies, merely that they should be considered. Some procedures may transfer directly from psychopharmacology to MCS, but others may need some modification, or may not be practicable at all for certain types of MCS phenomena.

There are two basic types of test designs. In the first, called a repeated-measures or within-subject design, subjects are given all treatments in counterbalanced orders. In psychopharmacology, the treatments would be drugs and placebos. In MCS, a repeated measures design could expose MCS sufferers to both a real chemical and to a sham, or placebo. It is important that the orders of exposure are balanced; half the subjects receive the real exposure first and the sham exposure second; the other half receive the sham exposure first and the real exposure second. Subjects should be allocated to the two orders at random. This balances the intercurrent effects such as familiarization with the procedures and learning on the tests themselves.

Repeated measures designs can be extended to cover more than two exposures, e.g., to measure sensitivity to more than one chemical. Balancing the orders of exposure is still important, and one way to do this is to cover all permutations of exposures; each subject follows one permutation. For example, if the effects of exposure to two chemicals and a sham as control are to be studied, there are three exposures in total: A, B, and C. There are six permutations of this: ABC, ACB, BCA, BAC, CAB, and CBA, and this means that at least six subjects must be tested to cover all permutations. This procedure can be replicated as many times as necessary to achieve a statistically viable number of subjects, with the total number of subjects being a multiple of six. The same random allocation procedures should be followed.

The same procedure can be followed for more than three exposures, but the number of permutations rises rapidly and quickly becomes unmanageable, e.g., there are 24 permutations of four exposures, which would require subjects in multiples of 24. Fortunately, there is a more elegant way of counterbalancing four exposures: a Latin Square. This is simply a way of selecting four of the permutations so that each exposure occurs once in each serial position. There are several Latin squares, but in psychopharmacology the squares are also chosen so that no given exposure precedes or follows another given exposure more than once. The square on the left, below, fulfills the first criterion but not the second; the square on the right fulfills both criteria.

| A | B | C | D |
|---|---|---|---|
| B | D | A | C |
| C | D | B | A |
| D | A | B | C |

Each subject follows a row chosen at random and it takes four subjects to complete the design. As before, enough replications are carried out to ensure a statistically viable number of subjects. In psychopharmacology, enough time is usually allowed between treatments for the drug to be completely metabolized before the next treatment is given. In MCS, more time would have to be allowed if the effects are longer lasting.

Latin squares can be devised to cover more than four exposures, but the process again quickly becomes unmanageable because it is often difficult to devise squares that are completely counterbalanced, and analysis and interpretation become cumbersome.

The advantages of repeated measures designs are that they usually require fewer subjects than other designs and the statistical variance is reduced, since the subjects are their own controls and do not have to be matched. Reducing the variance means that any differences between exposures can be more easily seen. The disadvantage is that the design is inflexible; if any subject misses an exposure or data are lost on one occasion through error or equipment breakdown, all of that subject's results from all of his or her exposures must be discarded.

In addition, using repeated measures designs, problems of asymmetric transfer and stimulus range effects (44, 45) may arise. Asymmetric transfer concerns unbalanced effects that could transfer between treatments. For example, an exposure might affect test performance differently depending on when it occurred in the treatment order, and this could affect performance on subsequent occasions. This could lead to over- or underestimation of the effects of the exposure. Stimulus range effects concern subjects' restricting or modifying their responses according to the range of stimuli they have experienced earlier in the experiment. It is a kind of learning effect. Some people consider these problems to be arcane and trivial, but others consider them to be so great as to render repeated measures designs unworkable (44, 45).

In the other main design, called a parallel groups or between-subjects design, subjects are divided into groups and each group receives one treatment. In psychopharmacology, the treatments would be drugs and placebos. In MCS, one group of patients could receive a real exposure to the chemical or agent to which they claimed to be sensitive and another group a sham exposure. This design can also be extended to cover any number of different chemicals simply by adding more groups. The groups do not have to contain the same numbers of subjects, although each group must contain at least the minimum for statistical analysis.

The advantage of parallel groups designs is that they are flexible; missing subjects or data are simply replaced and the studies can take less time. The disadvantages are that more subjects are required and there can be problems in selecting and matching appropriate control groups. If the groups are large, for example, more than 100 subjects, then allocation may be random. If the groups are small, as is more often the case, subjects must be matched. If they are not, differences between the groups could be confounded with differences between the treatments. Matching is more difficult than it seems, for the number of factors that could influence test performance is very large, and includes such things as physical and demographic factors, personality, intelligence, mood, background, knowledge, and experience. Some of these factors may appear obvious but have little effect on the results; some may appear unimportant, or may even be forgotten but might affect the results significantly. It is also possible that matching on one factor causes a mismatch on another.

In both types of design it is vital that neither the subjects nor the experimenters know which subjects have received which treatment. This is called a double-blind procedure and is intended to minimize the effects of bias, imagination, and prejudice. Subject allocation to treatment orders or groups should be carried out by a third party not involved in the experiment.

Other experiment designs are possible, but they are usually variations on, or hybrids of, the main two already described. Two variations are worth noting. The first is to use a parallel-groups design but test subjects before and after exposure. The results of these preexposure tests can be used to allocate subjects to exposure and control.
groups to ensure that they are matched in terms of their performance on the tests. Some experimenters consider this to be the most important matching criterion. Alternatively, preexposure tests can be used retrospectively to determine the degree to which the groups have been matched and to adjust the postexposure results accordingly.

The other variation, called a balanced-placebo design, is designed to separate the actual effects of a chemical from the anticipated or imagined effects to account for the possibility that MCS might be a conditioned response. In this design, there are four conditions: a) subjects are told they will be exposed and are actually exposed; b) subjects are told they will be exposed and are not exposed (or exposed to a placebo); c) subjects are told they will not be exposed and are not exposed; and d) subjects are told they will not be exposed and are exposed. These conditions are applicable to both repeated-measures and parallel-groups designs.

This design is similar to that followed in signal detection studies, and is used widely in applied psychology to separate ability to detect stimuli from bias in reporting it. Signal detection analysis is made in terms of hits (responding when a chemical is present), misses (not responding when a chemical is present), false alarms (responding when there is no chemical), and correct rejections (not responding when there is no chemical). This design might not work with all MCS phenomena, but it would work with some, e.g., caffeine. The mechanics of the design are feasible, but the ethics, particularly of the last condition, may prohibit its use.

In practice, the choice of design is often forced by circumstances rather than decided on theoretical grounds. For example, if the exposures needed in a repeated measures design take a long time and subjects cannot guarantee to devote that time, a repeated measures design may not be possible.

As many subjects as possible should be used in the tests to increase measurement precision and reduce variance. However, the law of diminishing returns applies, and as a general rule, it is not worthwhile to use more than about 30 subjects per group or subgroup if small sample statistics, e.g., t-tests, are intended. If more subjects are needed, then at least 100 should be used, and normal distribution statistics applied. Subgroups can be formed if there are enough subjects with sufficiently important characteristics, e.g., males and females; old and young; old, middle-age and young; married, separated, divorced, etc.

**Experiment Control**

Careful experiment design when using cognitive and psychomotor performance tests requires careful experiment control to avoid confounding effects of the drug or chemical with effects of other factors. The principle of experiment control is to eliminate any variable that could influence the results. If variables cannot be eliminated, they should be held constant or counterbalanced. If this cannot be achieved, the variables should be measured so allowances can be made. Some of the descriptions below are obvious; others may not be so obvious but are nonetheless important.

Subjects react to the tester’s approach and mood. Thus, the tester should always be the same person, have the same approach, and give a standardized set of instructions to every subject every time. Testers cannot afford to be happy, sad, bad-tempered, or to show any human attributes lest they affect the subjects differently on different occasions. Double-blinding helps.

Cognitive and psychomotor performance tests, by their very nature, are sensitive to changes in individual and environmental circumstances. Thus, testing should not be done just before or just after periods of stress such as illness, emotional upset, or physical exercise, or relaxation such as vacations or weekends unless the same conditions can be guaranteed every time. Most of the tests use visual stimuli, and most of those that do not, use auditory stimuli. Thus, if subjects wear spectacles, contact lenses, or hearing aids, they should wear them for all tests.

The testing environment can affect test performance and should be standardized in terms of ambient lighting, temperature, noise, and workstation ergonomics such as seating comfort, angle and height, computer displays and keyboards, and viewing and operating distances and angles.

The tests are also sensitive to circadian rhythms; thus, subjects should always be tested at the same time of day. Shift workers should be treated separately, since they will have made different adjustments to their body clocks, depending on the shift timings and how long they have been doing shift work. If circumstances cannot be controlled, they should be measured or at least noted, so they may be used as factors in analyzing the results.

Stopping subjects from smoking or from drinking tea, coffee, or alcohol may appear to give testers more control. This might be true in pharmacokinetic studies, in which the xanthesines (e.g., caffeine, theobromine, and theophylline), nicotine, or alcohol could interfere with absorption of the drug or chemical at issue, but in psychological studies this may be a mistake. First, smokers and alcohol drinkers, and even habitual coffee drinkers, will not stop just because they are asked to do so. Second, even if they do stop, the change in behavior pattern and the withdrawal symptoms could affect performance more than if they had continued with the nicotine, caffeine, or alcohol. Generally, it is better to let subjects follow their normal patterns of behavior, but it is wise to ask them how much of a particular substance they normally take and how much they have taken in the last 24 hr. Remember, xanthesines are also present in other foods such as cola beverages and chocolate.

Subjects learn with practice; the more they undergo tests, the more skilled they will become, even though the tests might involve well practiced skills. Learning varies with the tests; simple reaction time tests involve little learning, but tracking tests usually involve considerable learning. Learning tends to follow a negatively accelerating curve; it is greatest at the start and gradually reduces with repeated practice. The learning effect is very strong and often can have a greater impact than a drug or possibly a chemical.

Some psychologists spend their careers trying to improve learning, but performance psychologists often wish it did not exist, for it often interferes when they are trying to balance an experiment design. Some experimenters insist on extensive training beforehand so subjects can attain a plateau of performance that is then assumed to be stable throughout the experiment. Unfortunately, preexperiment training can be prohibitively protracted and performance plateaux are not what they seem; many are simply end-of-session effects. Any change between the training and experimental environment or test procedures will change subjects’ motivations and, hence, their performances and rates of learning. Thus, subjects may have reached a learning plateau at the end of training but start to improve when the experiment proper starts.

Nonetheless, it is important that subjects are at least guided through the first steep part of their learning curves. Three basic methods can be used: give all subjects the same amount of training regardless of how much they improve; train all subjects to the same criterion, e.g., to where their scores on successive tests are within, for
example, 10%, regardless of how long or how many tests this takes; or let them say when they have had enough and note how much they did. Most experimenters favor the first option because they are usually short of time, but the second option is better in that it allows for different learning rates. The third option is rarely used.

Sham exposures, or placebos, have been mentioned several times and some further comments on them are appropriate. The sham exposure should be the same as the real exposure in all respects except the real chemical is not used. For example, the same location, environment, and timings, including time of day, must be used; the same administration procedures must be followed and the same instructions given.

In psychopharmacology, placebos or sham drugs are relatively easy to administer double-blind; one simply gives an injection of isotonic saline or matches the size, shape, and color of any tablets given. If the tablets cannot be matched, they can be put in identical opaque gelatin capsules; this also has the advantage of hiding any taste or odor. In MCS, placebos can very difficult to devise because the taste and odor of the chemical cannot easily be removed. Alcohol is a case in point; it has a distinctive taste and odor, and many ingenious and imaginative means of disguise have been tried, including essences of rum or whisky, juniper (for gin), strong-tasting herbs and spices, and smearing the rim of the glass with alcohol. However, none has been completely satisfactory. Diluting the alcohol works, but the dilution must be so great that the quantities become prohibitive.

Smearing the rim of the glass with alcohol would not work in MCS because the taste and odor of the chemical can be the very factors that trigger the reaction rather than any pharmacological effect of the chemical itself. In MCS, making the placebo taste and smell like the test chemical would make the placebo active—it would not be a placebo any more. Masking the taste and odor of the test chemical with those of other chemicals could also be a problem because subjects could also react to the masking chemicals. It is characteristic of MCS that hypersensitivity, originally to one chemical, can later extend to other chemicals. It is even possible that the masking chemicals themselves could induce new hypersensitivity reactions.

Placebos are easier to devise if subjects know which environments induce reactions but are not aware of the taste or odor of the chemicals involved. For example, if subjects complain that they feel ill in certain buildings, rooms, or vehicles but do not know why, rooms or vehicles with the same appearance but without the taste or odor could be arranged. This might be difficult and expensive and in some cases impracticable, but clean rooms are used by some MCS investigators, and one should keep an open mind. If the effects of the components of the building or vehicle must be isolated, the components should be substituted as necessary without changing the overall appearance of the building or vehicle. For example, carpets and soft furnishings should not be added but replaced by items similar in all respects except that they contain the test chemical, or no chemical, as appropriate. Subjects would be bound to notice if a carpet were suddenly to appear on what had been a bare floor or a red carpet where there had been a blue one, and could react simply to its appearance rather than to the off-gassing chemicals.

In psychopharmacology, placebos are used routinely as control treatments, since they have no effects alone. However, it is sometimes necessary to use a treatment that does have effects, e.g., to add weight to a finding that the test treatments have no effect. Often, findings of no effect are taken to mean that there is no effect. This is wrong, however; absence of evidence is not evidence of absence. A treatment used to produce a known effect is called a verum (plural vera, from the Latin for truth).

A combination of vera and placebos could offer a means of avoiding the problem of the taste and odor of chemicals. If it is not possible to “blind” subjects by removing, disguising, or masking taste and odor, it may be possible to confuse them by exposing them to a succession of individual chemicals and mixtures with various tastes or odors, including placebos and vera. To aid the confusion, the number of exposures should exceed the subjects’ short-term memory spans; at least eight or nine exposures would be necessary. Vera for ingested chemicals could be drugs with known effects; vera for airborne chemical vapors are more difficult, but anesthetics might be suitable.

Double-blind procedures have been mentioned before, but one further comment is necessary. Double-blindling is widely accepted as a correct procedure during a study to avoid any bias caused by imagination and expectation. What is often forgotten is that bias can also occur during analysis of the results, especially if data are missing or need interpretation. Thus, double-blind procedures should continue until analysis is complete; experimenters do not need to know the identities of the exposures to analyze the results—they simply need labels.

In addition to cognitive and psychomotor performance tests, it may be necessary to include other measures to help interpret the results. The type of measure will vary with the type of study, but as a general rule, it would be useful to include histories of diet, smoking, and drinking xanthine- or alcohol-containing beverages. These measures could be taken in terms of behavior generally and behavior immediately before the study. Also, a large number of MCS sufferers show signs and symptoms of anxiety and depression, and it is not unreasonable to expect that many people would be anxious about an experimental exposure to an agent they suspect would exacerbate their problems. Thus, it would be useful to incorporate into the test standard measures of anxiety and depression together with a measure of mood to help explain any potential confounding of these effects with those of the agent itself. Last, because of the question of adaptation, it is important to include the exposure history of the MCS sufferers.

**Conclusion**

MCS is very real to the sufferer, whether the effects are physical or imagined, and whether or not others believe them. The effects can be serious, even debilitating, and the incidence is growing. Most of the symptoms can affect cognitive and psychomotor function, either directly or indirectly. Effects such as poor concentration and memory could be caused directly by an effect on the central nervous system, or indirectly through distraction by somatic effects. Somatic problems such as joint and muscle aches and irritation could also affect coordination directly. Thus, for whatever reason, an individual’s ability to think, reason, and coordinate could be impaired.

This paper has described several cognitive and psychomotor performance tests as well as some experiment designs and control procedures that could be considered for use in MCS studies. Cognitive and psychomotor performance tests have a record of long and successful use in psychopharmacology. The tests have been used to evaluate the effects of many drugs, both internally, i.e., that the drugs caused the effects, and externally, i.e., that the effects extrapolate to real life. Tests based on the stage processing
model of performance have also been used successfully to differentiate drug effects on various cognitive stages, e.g., discrimination, recognition, information storage, information recall, and decision making. Cognitive and psychomotor performance tests have also been widely used to assess the effects of other environmental stressors such as heat and cold, fatigue, wearing protective clothing, and exposure to industrial chemicals.

Cognitive and psychomotor performance tests could be put to the same use in MCS. They can measure the effects of experimental exposures to chemicals or other agents and determine the efficacy of diagnostic, preventative, or therapeutic measures. They can assess the ability of MCS sufferers to work, and they can help sway the opinion of those people who are sceptical of the authenticity of MCS by contributing to the body of objective evidence.

REFERENCES

1. Wetherell A. Performance tests. Environ Health Perspect 104:247–273 (1996).
2. AGARD. Human Performance Assessment Methods. AGARDograph 308. Neuilly-sur-Seine, France: NATO Advisory Group for Aerospace Research and Development, 1989.
3. Wetherell A. The STRES Battery: Standardised Tests for Research into Environmental Stress. In: Contemporary Ergonomics 1990 (Lovesey EJ, ed). London: Taylor and Francis, 1990:270–275.
4. Wesnes K, Simpson P, Christmas L. The assessment of human information processing abilities in psychopharmacology. In: Human Psychopharmacology: Measures and Methods (Hindmarsh I, Stonier PD, eds). Wiley: Chichester, 1987:79–91.
5. Hockey R, Hamilton P. The cognitive patterning of stress states. In: Stress and Fatigue in Human Performance (Hockey R, ed). Wiley: Chichester, 1983:331–362.
6. Parrott AC. The effects of transdermal scopolamine and four doses of oral scopolamine (0.15, 0.3, 0.6, 1.2mg) upon psychological performance. Psychopharmacol 89:347–354 (1986).
7. Holding DH. Skills research. In: Human Skills (Holding DH, ed). Wiley: Chichester, 1989:1–16.
8. Parrott AC. Performance tests in psychopharmacology 3: construct validity and test interpretation. Human Psychopharmacol 6:197–207 (1991).
9. Cull C, Trimble MR. Automated testing and psychopharmacology. In: Human Psychopharmacology: Materials and Methods. Vol 1 (Hindmarsh I, Stonier PD, eds). Chichester: Wiley, 1987.
10. Wittenborn JR. Psychomotor tests in psychopharmacology. In: Human Psychopharmacology: Materials and Methods. Vol 1 (Hindmarsh I, Stonier PD, eds). Chichester: Wiley, 1987.
11. Hindmarsh I. Psychomotor function and psychoactive drugs. Br J Clin Pharmacol 10:189–209 (1980).
12. Smith PJ, Langolf GD. The use of Sternberg’s memory scanning paradigm in assessing the effects of chemical exposure. Human Factors 23:701–708 (1981).
13. Maizlish NA, Langolf GD, Whitehead LW, Fine LJ, Albers JW, Goldberg J, Smith P. Behavioural evaluation of workers exposed to mixtures of organic solvents. Br J Ind Med 42:579–590 (1985).
14. Holland P, Kemp KH, Wetherell A. Some effects of 2 mg atropine and 5 mg diazepam, separately and combined, on human performance. Br J Clin Pharmacol 5:367P (1978).
15. Wetherell A. Some effects of atropine on short-term memory. Br J Clin Pharmacol 10:627–628 (1980).
16. Toulmin SJ, Wetherell A. Some effects of anticholinergic drugs on performance. In: Contemporary Ergonomics, 1995 (Lovesey EJ, ed). London: Taylor and Francis, 1995.
17. Wetherell A. Performance effects of physostigmine and scopolamine as nerve agent pretreatments. Proc Medical Defence Bioscience Review. Vol 2. U.S. Army Medical Research and Development Command, 1995:653–661.
18. Wetherell A. Drug and driver’s visual perception. In: Vision in Vehicles (Gale AG, Freeman MH, Smith P, Taylor SP, eds). Amsterdam: North Holland, 1986:33–42.
19. Wetherell A. Some factors affecting spatial memory for route information. In: Information Design (Easterby RS, Zwaga HJG, eds). Chichester: Wiley, 1984:321–338.
20. Wetherell A. Effects of physostigmine on stimulus encoding in a memory scanning task. Psychopharmacologia 109:198–202 (1992).
21. Shattuck JA, Wetherell A. Cognitive and psychomotor performance during treatment with ondansetron and granisetron. J Psychopharmacol 9(Suppl): A40 (1995).
22. Wetherell A, Shattuck JA, Cook JM. The ergonomics of nuclear, biological and chemical individual protective equipment. In: Contemporary Ergonomics, 1992 (Lovesey EJ, ed). London: Taylor and Francis, 1992.
23. Shingledecker CA. A task battery for applied human performance assessment research. Report AFAMRL-TR-84-071. Dayton, OH: Wright-Patterson Air Force Base, 1984.
24. Moran L, Mefferd RB. Repetitive psychometric measures. Psychol Rep 5:269–275 (1959).
25. Baddeley AD. A 3-minute reasoning test based on grammatical transformation. Psychon Sci 10:341–342 (1968).
26. Lewis VJ, Baddeley AD. Cognitive performance, sleep quality and mood during deep oxyhelium diving. Ergonomics 24:773–793 (1981).
27. Logic RH, Baddeley AD. A trimix saturation dive to 660 m: studies of cognitive performance, mood and sleep quality. Ergonomics 26:359–374 (1983).
28. Fitts PM, Weinstein M, Rappaport M, Anderson N, Leonard JA. Stimulus correlates of visual pattern perception: a probability approach. J Exp Psychol 51:1–11 (1956).
29. Chiles WD, Alluvi EA, Adams OS. Work Schedules and performance during confinement. Hum Factors 10:143–196 (1968).
30. Jex HR, McDonell JD, Phatak AV. A "critical" tracking task for manual control research. IEEE Transactions on Human Factors in Electronics HFE-7, 1966:138–144.
31. Boer LC, Gaillard AWK, Jorna PGAM. Taskomat, a task battery for information processing. Rpt IZF 1982-2. Soesterberg, The Netherlands: Institute for Perception TNO, 1982.
32. Leonard JA. Five choice serial reaction apparatus. Rpt 326/59. Cambridge, MA: Medical Research Council Applied Psychology Unit, 1959.
33. Broadbent DE, Broadbent MHP, Jones JL. Performance correlates of self-reported cognitive failure and obsessionality. Br J Clin Psychol 25:285–299 (1986).
34. Wesnes K, Warburton DM. Smoking, nicotine and human performance. Pharmacol Ther 21:189–208 (1983).
35. Stroop JR. Studies of interference in serial verbal reactions. J Exp Psychol 18: 643–662 (1935).
36. Michon JA. Tapping regularity as a measure of perceptual motor load. Ergonomics 9:401–412 (1966).
37. Baddeley AD. The capacity for generating information by randomisation. Quart J Exp Psychol 18:119–129 (1966).
38. Wechsler D. A standardised memory scale for clinical use. J Consult Clin Psychol 19:87–95 (1955).
39. Sternberg S. High-speed scanning in human memory. Science 153:652–654 (1966).
40. Sternberg S. Two operations in character recognition: some evidence from reaction time measurements. Percept Psychophys 2:45–53 (1967).
41. Sternberg S. Memory scanning: mental processes revealed by reaction time experiments. Am Sci 57:421–457 (1969).
42. Sternberg S. The discovery of processing stages: extensions of Donders' method. Acta Psychol Attention Performance 30:276–315 (1969).

43. Massaro DW. Experimental psychology and information processing. Chicago: Rand McNally, 1975.
44. Poulton EC, Freeman PR. Unwanted asymmetrical transfer effects with balanced experimental designs. Psychol Bull 66:1–8 (1966).
45. Poulton EC. Range effects in experiments on people. Am J Psychol 88:3–32 (1975).