CONFERENCE REPORT

Role of human neurobehavioural tests in regulatory activity on chemicals

Richard Stephens, Penny Barker

Abstract

Psychological performance tests have been used since the mid-1960s in occupational and environmental health toxicology. The interpretation of significantly different test scores in neurobehavioural studies is not straightforward in the regulation of chemicals. This paper sets out some issues which emerged from discussions at an international workshop, organised by the United Kingdom Health and Safety Executive (HSE), to discuss differences in interpretation of human neurobehavioural test data in regulatory risk assessments. The difficulties encountered by regulators confronted with neurobehavioural studies seem to be twofold; some studies lack scientific rigor; other studies, although scientifically sound, are problematic because it is not clear what interpretation to place on the results. Issues relating to each of these points are discussed. Next, scenarios within which to consider the outcomes of neurobehavioural studies are presented. Finally, conclusions and recommendations for further work are put forward.

Keywords: neurobehavioural tests; regulation of chemicals; interpretation

Ergonomics and Work Psychology Section, Health and Safety Laboratory, Broad Lane, Sheffield, UK
R Stephens

Health and Safety Executive, Health Directorate D3, Room 146, Magdalen House, Stanley Precinct, Bootle, Merseyside, UK
P Barker

Correspondence to:
Dr R Stephens, Ergonomics and Work Psychology Section, Health and Safety Laboratory, Broad Lane, Sheffield, S3 7HQ, UK.

Accepted 4 October 1997

Psychological performance tests have been used since the mid-1960s in occupational and environmental health toxicology.¹ Within this context they are usually referred to as neurobehavioural (or neuropsychological) tests. Such tests have been used to investigate whether effects on the nervous system occur after acute² or long term³ occupational exposures to substances, typically metals—including lead,⁴ manganese,⁵ mercury,⁶ and aluminium⁷—solvents—such as styrene, tetrachloroethylene, toluene, methyl ethyl ketone, trichloroethane, trichloroethylene, white spirit, and xylene⁸—and pesticides—organophosphates,⁹ and carbamates.¹⁰ Neurobehavioural tests measure effects on the nervous system through simple tasks of motor speed and coordination, as well as through more complex memory and attention tasks.¹¹ In a typical study, several neurobehavioural tests are given at one time, to a group of people, as a test battery.¹² The gen-

erally accepted definition of a neurobehavioural effect is that there is a significant difference in the score on a neurobehavioural test of a group before and after exposure, between exposed and control groups, or between groups exposed at different levels.

There exists an extensive international framework of chemical regulation designed to prevent or minimise effects of exposure on human health. Although there is considerable conformity on some ill health effects where—for example, when the Organisation for Economic Co-operation and Development (OECD) test methods¹³ and the European Union criteria for classification for hazard¹⁴ are used—for neurobehavioural effects wide divergence persists in interpretation. For instance, many more patients are reported to have been diagnosed as having toxic encephalopathy (the diagnostic criteria for which require neuropsychological test deficits) in Denmark compared with other industrialised European countries.¹⁵

Problems with the interpretation of results from neurobehavioural studies are exemplified by the case of neurobehavioural studies on 1,1,1-trichloroethane, reported in the published review *Environmental Health Criteria* (EHC) 136.¹⁶ The critical effect in humans is central nervous system depression. Several neurobehavioural volunteer studies were available. The task group concluded that the no effect level from several of these studies seems to be in the region of 250 ppm (parts per million). It was noted that one study suggested effects at a lower level (175 ppm) but “the interpretation of the results was unclear.” This particular study reported slowed reaction time but improved attention and concentration (and no change across other test outcomes).¹⁷ It seems that the task group for the EHC was unable to reach a firm consensus on this study. By contrast with some regulatory scientists elsewhere, technical advisers in the Health and Safety Executive (HSE) regard the no adverse effect level to be 250 ppm and do not attach any weight to the results at 175 ppm in occupational regulatory activity on 1,1,1-trichloroethane.

The HSE held an international workshop to discuss differences in interpretation of human neurobehavioural test data in regulatory risk assessments. Participants included both

Table 1 Factors that potentially confound or modify

Stable factors	Varying factors
Age (date of birth)	Alcohol (recent use)
Educational level	Caffeine (recent use)
Sex	Nicotine (recent use)
Socioeconomic group or occupation	Medicines or drugs (recent use)
First language	Paints, glues, or pesticides (recent use)
Preferred hand	Near visual acuity
Computer experience (automated tests only)	Upper body injury restricting movement
Caffeine (habitual use)	Recent cold or flu
Alcohol (habitual use)	Stress
Nicotine (habitual use)	Arousal (and fatigue), sleep last night
Medicines or drugs (habitual use)	Screen luminance (automated tests only)
Paints, glues, or pesticides (habitual use)	Time of day
Diabetic	Time of year
Epileptic	
Other nervous system disease	
Head injury causing unconsciousness for > 1 h	
Alcohol or drug addiction problem	
Level of physical activity at work	

regulatory toxicologists and neurobehavioural toxicology researchers. This paper sets out some issues which emerged from discussions at the workshop, with the aims of both encouraging debate in neurobehavioural toxicology, and of promoting some aspects relating to good practice, so that neurobehavioural toxicology studies may better be incorporated into the regulatory risk assessment process.

PROCESS OF ASSESSMENT

The difficulties encountered by regulators confronted with neurobehavioural studies seem to be twofold. Some studies lack scientific rigor; put more precisely, they contain threats to internal and external validity.¹⁸ Other studies, although scientifically sound, are problematic because it is not clear what the implications of their reported effects are. This suggests that a two step strategy is appropriate for evaluating neurobehavioural studies in the context of regulation; such an approach was advocated by a similar initiative recently concluded in Germany. This involves a process of first assessing study methods, to determine whether an effect has been rigorously shown to be related to a specific exposure with epidemiological and other scientific criteria. Only after this is the second step considered of interpreting the meaning of that effect with respect to regulation. Issues raised at the workshop are discussed within this two step strategy in this paper.

First step: has a neurobehavioural effect been shown?

STUDY DESIGN

The following points may be useful to those designing and evaluating studies.

Selection bias

Predetermined criteria for selection of a study group should be applied during recruitment of exposed and control groups; response and drop out rates can indicate potential bias, and should be reported.

Separation of acute and chronic effects

In studies of possible chronic effects, it should be clear that an interval, sufficient to allow any acute exposure effects to subside, has elapsed between the most recent exposure and testing.

If such an interval is not allowed, then a false positive chronic effect may result, which is, in fact, an acute exposure effect.

Confounding or modifying factors

The table provides a list of a large range of factors that can influence human performance. As a minimum, studies should control for effects of age, sex, educational level, and socioeconomic level. This may be achieved through matching of the control group (possibly excluding potential participants with low frequency characteristics). Statistical procedures are often used to adjust for the influence of possible additional confounders or modifiers. Repeated measures designs have the advantage of restricting covariance by making each participant their own control, although practice effects may add confusion. Practice effects may be controlled with counter balanced study designs, careful selection of tests more resistant to practice effects, or training of participants to an optimal level of performance before baseline testing. Ambient conditions, such as temperature and background noise, should be kept constant as far as possible. Although, historically, studies have attempted to control for premorbid ability with tests of so called "hold" abilities—such as vocabulary tests—this procedure has been questioned recently.¹⁹

Tester reliability

For studies that use non-automated tests, measures to control reliability within and between testers—such as blinding—should be incorporated and reported.

Exposure data

Regulatory toxicology relies on determining exposure levels at which effects do and do not occur. Hence it is imperative that positive neurobehavioural studies define clear dose-response relations. The (often poor) quality of exposure data was a major criticism of neurobehavioural studies at the workshop.

LITERATURE REVIEWS

When assessing a series of studies, the Bradford-Hill²⁰ or similar criteria should be applied to determine causality. We discussed several of these criteria with particular reference to neurobehavioural studies.

Consistency and specificity

Replication of an observed effect across different studies conducted by different investigators makes for a powerful argument that the effect is genuine. The trend in all the studies that have been conducted is one consideration in reaching a conclusion about the effect being real. An unresolved issue in neurobehavioural toxicology, however, is whether replication should be sought simply at test level, or at the level of the function which tests measure. Test level replication would look for similar decrements in performance on the same test across studies. Alternatively, different studies might use different tests which measure a common function,

such as short term memory. Decrements in these tests across studies would be considered to be replication at the functional level.

Replication is sought at test level because the functions underlying the tests favoured in neurobehavioural studies are not well understood. These tests were developed for purposes other than toxicology, such as measuring intelligence, but their continued use was encouraged by initial positive findings.²¹ By developing standardised neurobehavioural test batteries researchers eliminated the necessity to know what a test was measuring when seeking replication. Instead, standardisation encourages the use of identical tests across studies, and replication across tests is sought. Standard test batteries include the consensus recommended World Health Organisation neurobehavioural core test battery,²² the computerised neurobehavioural evaluation system²³ and the Swedish performance evaluation system.²⁴ These batteries each contain similar collections of traditional tests.

This approach of standardising on tests which measure functions that are not clearly defined has certain disadvantages. One disadvantage is that standardisation necessarily and actively discourages the development of new and improved methods. This could be viewed as a premature halt in the development of what is, in any case, an immature discipline (only a small fraction of the thousands of industrial chemicals in use have been investigated with neurobehavioural methods). A second disadvantage with standardisation is the difficulty in assessing the significance (in a regulatory setting, for example) of decrements in tests the underlying functions of which are not clearly defined. Thus some newer neurobehavioural test batteries, particularly the automated cognitive testing system²⁵ and the information processing and performance test battery,²⁶ have adopted a different approach based on cognitive psychology theory. Derived from research into normal intellectual functioning, these focus more on the psychological functions which underlie performance of a test. However, these batteries are currently less widely used than the standardised batteries.

Coherence

The requirement for coherence across different types of data (biochemical, animal, and human) emerged as being a main criterion required for human neurobehavioural data to be incorporated into regulatory decision making in the United Kingdom. Some researchers argued that they would place less emphasis on this criterion if an acceptable level of consistency or specificity of the effect of a substance had been shown.

Mechanistic plausibility

It may be argued that the existing base of neurobehavioural studies is not large enough, currently, to facilitate sufficient understanding of biologically plausible underlying mechanisms.²⁷ Thus some researchers argued that they would be prepared to accept a repeatedly

demonstrable neurobehavioural effect in the absence of such a mechanism.

Overall, studies should be individually scrutinised for the quality of study design and reporting, and conclusions about whether there is evidence for an exposure-related response should be drawn from consideration of all the available studies, attaching more weight to those of better quality. It might be possible to conclude that the substance causes a significant neurobehavioural effect at a particular exposure concentration. The potential consequence of that effect must then be considered.

Step two—interpretation of neurobehavioural effects in the regulation of chemicals

ASSESSING THE NATURE AND SEVERITY OF NEUROBEHAVIOURAL EFFECTS

As already discussed, part of the problem faced when assessing the nature of neurobehavioural effects arises out of the lack of understanding of the functions that neurobehavioural tests measure, and therefore, what changes in those functions mean. This can be rectified by adopting a more theoretical psychological approach, which, as already noted, some newer test batteries have done. Applying contexts to neurobehavioural effects, through benchmarking or calibrating techniques, can also facilitate understanding of their nature and severity. Such techniques involve comparing the effects of an occupational exposure with the effects of a reference substance or condition—such as alcohol or aging—the effects of which are better known to those receiving the study. In this way, both the nature and the severity are illustrated by analogy. One problem associated with this approach is that the specificity of effects of an exposure may be similar in some ways, but different in others to those of the comparison agent. Difficulties also arise—for example, in alcohol intoxication—translating the kinds of chronic effects found in epidemiological studies into acute effects of a reference substance. A further consideration is that errors in outcome measurements or exposure classification often occur in epidemiological studies,²⁸ and these may introduce null bias. Null bias increases the probability of false negative studies, and attenuates significant findings,²⁹ hence decreasing study sensitivity. The possibility that measured neurobehavioural effects are underestimates of true effects, owing to null bias, should also be considered when assessing the severity of neurobehavioural effects found in epidemiological studies.

DEFINING CRITERIA FOR ADVERSE EFFECTS

As well as appreciating the nature and severity of a neurobehavioural effect, its likely consequences, in terms of present and future risk from perspectives of health, safety, or other factors, need to be considered when deciding if it should be treated as adverse.

When effects are transitory—such as acute effects found in volunteer exposure studies—then immediate consequences, including safety or fitness to work issues, are one context in

which a regulatory decision could be made. A second consideration might be whether such effects are viewed as health effects. Transitory effects might also have an economic or public safety consequence. For instance, the economic consequence of the risk from sick building syndrome may be very large due to the large exposed population (potentially, all office workers). Also the public health consequence of an operator's failure to maintain vigilance in a situation where safety is critical—such as in a nuclear reactor control room—might be large enough for the transitory effect to be deemed adverse.

Prevalent in the neurobehavioural literature are studies reporting subclinical neurobehavioural effects of chronic (low level) exposure. By definition these effects may not be perceptible to study participants. Indeed, such exposure effects may be smaller than similar effects of established performance modifying factors—for example, age, sex, educational level, diurnal rhythms, sufficiency of sleep, and the effects of alcohol and drugs. Regulators questioned how such effects, likely to be imperceptible to a person, and in any case within the range of background variability in the population, can be considered adverse. Several arguments as to how such effects might be considered to be adverse were advanced.

One argument was that as neurobehavioural tests are complex tertiary measures, with no understood mechanism, any change in the direction of poorer performance should be assumed to be adverse, until there is evidence to the contrary. However, few would agree that neurobehavioural effects should, almost by definition, be considered to be adverse, as this argument requires. Neither does this occur in practice, as there are data indicating effects on colour discrimination with exposure to 11 ppm styrene,³⁰ but no countries regulate styrene to that level. A second argument was that although the consequences of a significant, although relatively small, effect would be imperceptible to most people, effects would be perceptible to those whose scores fall in the tails of the distributions of population scores and therefore could be considered to be adverse.

A third argument was that a relatively small downward shift in a variable—such as intelligence quotient (IQ)—could have an economic impact if the exposed population is very large. For example, the regulatory decision to remove lead from petrol in the United States was based on data showing an inverse relation between intelligence test scores and lead exposure, in groups of children, because of the economic implications to the country as a whole of a mass downward shifting in IQ. This was despite the groups studied scoring above average on the test used.³¹

A fourth argument was that, with continued exposure, currently imperceptible but measurable deficits will gradually become perceptible as health effects. The notion of subclinical neurobehavioural decrements providing early warning is much cited in the neurobehavioural literature.^{11–32} However, there exists little or no

evidence in support of such a continuum between subclinical neurobehavioural effects and clinical disease.

Conclusions

This article has outlined scenarios in which neurobehavioural effects might be considered to be adverse. No specific criteria are offered here to define what is adverse. This is because it is our opinion that the studies being conducted, and the tests being used, are currently too diverse to be able to set such criteria. Rather, neurobehavioural data will need to be considered on a case by case basis to determine whether regulatory action should be taken for particular substances. However, the workshop has emphasised the need for researchers to tailor neurobehavioural studies, both in design (particularly improving exposure data) and reporting, to the requirements of regulatory toxicology. By the same token, increased knowledge of psychology testing would assist those involved in the process of interpretation of neurobehavioural data for the regulation of chemicals.

Some areas for research are suggested. The standardisation effort in neurobehavioural studies may require review, as the tests currently promoted have limitations. One limitation of the standard tests is that the mental functions which underlie them are often poorly understood, and decrements may be referred to in publications by misleading terms such as “visual intelligence” or “logical memory”, despite there being no evidence that such mental functions exist.⁸ Newer tests specifically developed to measure functions defined by contemporary psychology research—for example, “working memory”—may yield more interpretable results, as they measure functions for which there is a knowledge base. This knowledge would be useful to regulators in determining the potential consequences of deficits in such functions. Existing data need not go to waste; bench marking the effects measured by well established tests would improve their interpretability, as would investigating further their underlying functions. The area of subclinical neurobehavioural effects would also benefit from further research, particularly examining whether there is a link with any future health effects. Although the best method of researching this question would be to conduct prospective studies, a quicker answer might be provided by the approach used by Echeverria *et al.*,³³ in which the results of neurobehavioural tests conducted with asymptomatic occupational groups are compared with results on the same tests in patients with suspected exposure related clinical nervous system disease.

In the meantime, regulatory activity continues against a background of insufficient information about clinical outcomes. To some extent, however, the same applies to the interpretation of certain physiological variables, such as lung function, in which a deficit occurs in the absence of clinical illness. The HSE is therefore looking at parallels that can be drawn between neurobehavioural and

physiological test models. The changing public perception of the concept of health, from a simple disease free concept, to quality of life issues could lead to more rigorous debate about the significance of effects such as those found in neurobehavioural studies.

This paper follows on from the discussions recorded at a workshop sponsored by the Health and Safety Executive on 8–9 February 1996.

- 1 Anger WK, Cassitto MG, Liang YX, *et al.* Comparison of performance from three continents on the WHO recommended neurobehavioural core test battery. *Environ Res* 1993;**62**:125–47.
- 2 Stollery BT, Flindt MLH. Memory sequelae of solvent intoxication. *Scand J Work Environ Health* 1988;**14**:45–8.
- 3 Stephens R, Spurgeon A, Calvert IA, *et al.* Neuropsychological effects of long-term exposure to organophosphates in sheep dip. *Lancet* 1995;**345**:1135–9.
- 4 Spurgeon A. Occupational lead exposure: do current exposure standards protect workers from harm? A review of neurobehavioural findings since 1978. *Indoor Environ* 1994;**3**:112–8.
- 5 Lucchini R, Selis L, Folli D, *et al.* Neurobehavioural effects of manganese in workers from a ferroalloy plant after temporary cessation of exposure. *Scand J Work Environ Health* 1995;**21**:143–9.
- 6 Kishi R, Doi R, Fukuchi Y, *et al.* Residual neurobehavioural effects associated with chronic exposure to mercury vapours. *Occup Environ Med* 1990;**51**:35–41.
- 7 Sjögren B, Iregren A, Frech W, *et al.* Effects on the nervous system among welders exposed to aluminium and manganese. *Occup Environ Med* 1996;**53**:32–41.
- 8 Gamberale F. Use of behavioural performance tests in the assessment of solvent toxicity. *Scand J Work Environ Health* 1985;**11**(suppl 1):65–74.
- 9 Rosenstock L, Keifer M, Daniell WE, *et al.* for the Pesticide Health Effects Study Group. Chronic central nervous system effects of acute organophosphate intoxication. *Lancet* 1991;**338**:223–7.
- 10 Verberk MM, Brouwer DH, Brouwer EJ, *et al.* Health effects of pesticides in the flower-bulb culture in Holland. *Med Lav* 1990;**81**:530–41.
- 11 Hartman DE. *Neuropsychological toxicology. Identification and assessment of human neurotoxic syndromes*, 2nd ed. New York: Plenum Press, 1995.
- 12 Spurgeon A. The use of neurobehavioural test batteries for research, diagnosis and screening: methodological aspects. *Toxicol Lett* 1995;**82–83**:191–5.
- 13 ECD. *OECD Guidelines for testing of chemicals*. Paris: OECD, 1993.
- 14 Commission Directive 93/21/EEC. *Official Journal of the European Communities L 110 A* 1993;**36**.
- 15 Triebig G, Bleecker M, Gilioli R, *et al.* International working group on the epidemiology of the chronic neurobehavioural effects of organic solvents: a need for further research! *Int Arch Occup Environ Health* 1989;**61**:423–4.
- 16 IPCS. *Environmental Health Criteria 136. 1,1,1-trichloroethane*. Geneva: World Health Organisation, 1992.
- 17 Mackay CJ, Campbell L, Samuel AM, *et al.* Behavioral changes during exposure to 1,1,1-trichloroethane: time-course and relationship to blood solvent levels. *Am J Ind Med* 1987;**11**:223–9.
- 18 Campbell DT, Stanley JC. *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally College, 1966.
- 19 Michélsén H, Lundberg I. Neuropsychological verbal tests may lack "hold" properties in occupational studies of neurotoxic effects. *Occup Environ Med* 1996;**53**:478–83.
- 20 Hill AB. The environment and disease: association or causation? *Proc R Soc Med* 1965;**58**:295–300.
- 21 Beaumont, JG. Neurobehavioral tests: problems, potential, and prospects. In: Russell RW, Flattau PE, Pope AM, eds. *Behavioral measures of neurotoxicity: report of a symposium*. Washington DC: National Academy Press 1990:86–97.
- 22 Johnson BL, ed. *Prevention of neurotoxic illness in working populations*. Chichester: John Wiley, 1987.
- 23 Baker EL, Letz R, Fidler AT. A computer-administered neurobehavioural evaluation system for occupational and environmental epidemiology. Rationale, methodology and pilot study results. *J Occup Med* 1985;**27**:206–12.
- 24 Iregren A, Gamberale F, Kjellberg A. SPES: a psychological test system to diagnose environmental hazards. *Neurotoxicol Teratol* 1996;**18**:485–91.
- 25 Stollery B. Cognitive assessment of toxicity: the automated cognitive test (ACT) system. In: *Computer-aided tests of drug effectiveness. Proceedings of a colloquium organised by the Institution of Electrical Engineers*. London: IEE, 1994.
- 26 Williamson AM. The development of a neurobehavioral test battery for use in hazard evaluations in occupational settings. Special issue: methods in behavioral toxicology and teratology. *Neurotoxicol Teratol* 1990;**12**:509–14.
- 27 Cory-Slechta DA. Bridging experimental animal and human behavioral toxicology studies. In: Russell RW, Flattau PE, Pope AM, eds. *Behavioral measures of neurotoxicity: report of a symposium*. Washington DC: National Academy Press 1990:137–58.
- 28 Rothman KJ. *Modern epidemiology*. Boston: Little, Brown, 1986.
- 29 Hernberg S. *Introduction to occupational epidemiology*. Michigan: Lewis, 1992.
- 30 Chia SE, Jeyaratnam J, Ong C, *et al.* Impairment of color vision among workers exposed to low concentrations of styrene. *Am J Ind Med* 1994;**26**:481–8.
- 31 Weiss B. The scope and promise of behavioral toxicology. In: Russell RW, Flattau PE, Pope AM, eds. *Behavioral measures of neurotoxicity: report of a symposium*. Washington DC: National Academy Press 1990:395–414.
- 32 Gamberale F, Kjellberg A, Akerstedt T, *et al.* Behavioral and psychological effects of the physical work environment: research strategies and measurement methods. *Scand J Work Environ Health* 1990;**16**(suppl 1):5–16.
- 33 Echeverria D, White RF, Sampaio C. A behavioral evaluation of PCE exposure in patients and dry cleaners: a possible relationship between clinical and preclinical effects. *J Occup Environ Med* 1995;**37**:667–80.

Occupational and Environmental Medicine and the electronic age

OEM has an Email address which is 100632.3615@compuserve.com. We welcome contact by Email, including letters to the editor. Many of our reviewers send us their reports by Email, helping to speed up the peer review process.

Our publishing system is now fully electronic, and authors are sending their revised copy to us on disk as well as paper. Watch for revised Instructions to Authors.

The Editor