III Safety Evaluation of Natural and Synthetic Flavourings

By K. R. Butterworth
THE BRITISH INDUSTRIAL BIOLOGICAL RESEARCH ASSOCIATION, WOODMANSTERNE ROAD, CARSHALTON, SURREY SM5 4DS

It is well known that the number of flavouring substances which need to be evaluated for safety is large, and flavours are only a small, albeit important, part of the group of chemicals which are employed in modern food technology. In recent years several official reports have been published which classify these natural and artificial flavouring substances, for example, that of the Council of Europe (1974)[40] and last year the Food Additives and Contaminants Committee (FACC) Report (1976).[41]

Table 8 *Mode of listing in Council of Europe document on natural and artificial flavouring substances* (1974)

Category	*Description of artificial flavouring substances*	*No. of materials included in class*
I (1—2000)	Those which may be added to foodstuffs without hazard to public health	692
II (2001—4000)	Those which may be added *temporarily* to foodstuffs without hazard to public health	293
III (4001—6000)	Those which are not admissible at present because technological and toxicological data are absent	243
IV (6001—)	Those which are not admissible since biological data indicate toxicity	3

In the 1974 Council of Europe 'Blue Book' the artificial flavouring substances were classified as shown in Table 8. A number was allocated to each material according to the group into which it fell. When the FACC report (1976) came out the 6000 series was omitted, and now it is expected that a new Council of Europe report will appear soon from which the members of both the 4000 and the 6000 series will be omitted. Even the materials in the 1—2000 series are not sacrosanct as the FACC report (1976) shows, since it requires further studies on several of these flavourings (*e.g.* propenylguaethol).

[40] Natural Flavouring Substances, their Sources, and Added Artificial Flavouring Substances. Council of Europe, Strasbourg, 1974.
[41] Food Additives and Contaminants Committee (1976). Report on the Review of Flavourings in Food. HMSO, London, 1976.

Table 9 *Studies required on temporarily admissible synthetic flavours* (2001—4000) *to investigate the possibility of upgrading members to the group which may be added to foodstuffs without hazard to public health* (1—2000)

13	Long-term studies
151	90-day studies
24	Metabolic studies
92	Hydrolysis studies
66	Acute studies

The figures in Table 9 are based on the original 'Blue Book' and show the number of tests required to upgrade the second group of 293 temporarily admissible artificial flavourings to the first group, where it is considered that they may be added to foodstuffs without hazard. The actual details in Table 9 are not important, but what is interesting is that an estimate of the cost to undertake these studies is £3.5 million at present day prices, and if the studies were carried out only at one set of laboratories which are equipped to undertake the work, such as BIBRA, they would take something like 35 years to complete. It is stressed that this is only to investigate the possible upgrading of members of the second group of synthetic flavourings to that of the first, which may be added to foodstuffs without hazard to public health. This list does not include the natural or the nature-identical flavours which have been used for many years and therefore are said to be less likely to have serious harmful effects in man.

The testing of flavourings is no different from any other food additive or contaminant. First of all it is necessary to ensure that the material to be investigated has an adequate specification. This has to be realistic, *i.e.* not so impure that the results of the subsequent toxicity studies are meaningless, and not so pure that it would be too expensive to produce this flavouring for industrial use. Tests performed on an unidentifiable material are completely worthless. Detailed analysis of all the constituents present in the material is essential and, in this context, technical specifications are generally inadequate because they are written for the purpose of selling the material.

To check the composition and purity of a material, any of the facilities of a well equipped chemistry laboratory may be required. These can range from simple measurements like melting points to elaborate examinations like mass spectroscopy. A good example is Strawberry Aldehyde, one of three flavourings which are in the Council of Europe (1974) 'Blue Book' in the 6000+ series, and thus not recommended for use in foods. It is likely that this is because this material has been reported to cause paralysis of the hind limbs of rats.[42] Samples of Strawberry Aldehyde were obtained from five different manufacturing houses and it was found that on gas–liquid chromatographic examination between two and seven peaks could be found in these materials (Table 10). The two peaks were believed to be due to the *cis* and *trans* isomers of the parent material,

[42] F. Griepentrog, *Med. Ernähr.*, 1969, **10**, 89.

ethyl methylphenyl glycidate, and the other peaks to contaminants accidentally or deliberately added.[43] In long-term studies in rats and other species, Mason and co-workers[44] have not been able to detect any neurological disorders and it is probable that the toxic effects reported previously were due to a contaminant rather than to ethyl methylphenyl glycidate itself. It is to be hoped that this important strawberry flavour may soon be upgraded.

Table 10 *Gas–liquid chromatographic finger-prints of 'Strawberry Aldehyde' samples*

		Area (%) of peak number						
		1	2	3	4	5	6	7
Supplier	*Retention time/min*	1.43	4.00	4.42	5.87	6.53	9.00	9.92
A		1.3	1.6	31.5	2.5	33.0	0.9	28.3
B		0.3	0.1	39.8	0.1	57.6	0	2.1
C		0.3	0	39.6	0.1	58.0	0	2.0
D		0.2	0	33.5	0.1	57.2	0	9.0
E		0.1	0.2	40.5	0.1	58.2	0	0.9

Most people know what an LD_{50} value is, and therefore it is only necessary to define it and outline its limitations. 'An LD_{50} is the dose which will kill 50% of a group of animals in a specified time (*e.g.* 24 hours or 7 days)'. Nowadays the limitations of such an LD_{50} test are obvious. For example, usually the material is administered on one occasion to each animal and then the animal observed for only a few hours or, at the most, days. Also, the chief concern is to determine how many animals die, rather than finding out the cause of death and attempting to detect any effects less dramatic than death. Occasionally attempts are made to assess the cause of death by simple observations. These include noting whether the respiratory movements cease before the heart stops, or whether the animal convulses and, if so, what is the type of convulsion. However, even today, it is rare to perform a post-mortem examination of animals in such an LD_{50} test. The importance of this point can be illustrated by an animal that might have died from, for example, a perforated stomach, which would have caused internal haemorrhage. This in its turn might have shown up as the heart stopping before the respiratory movements. Thus it might have been concluded that this was an effect primarily on the cardio–vascular system which, of course, would have been wrong.

Nowadays the LD_{50} test has a very limited place in the assessment of toxicity and a much more meaningful assessment of safety of flavourings can be made from short- and long-term studies. The 90-day test in the rat is a typical example

[43] P. L. Mason, K. R. Butterworth, I. F. Gaunt, P. Grasso, and S. D. Gangolli, *Food Cosmet. Toxicol.*, in the press.

[44] P. L. Mason, K. R. Butterworth, I. F. Gaunt, P. Grasso, and S. D. Gangolli, report in preparation.

of such a short-term study. In a typical study of this duration four groups, each of 25 male and 25 female weanling rats are employed. One group acts as a control and the others receive low, medium, or high doses of the materials.

Figure 4 shows the routine in such a study. At two and six weeks, interim

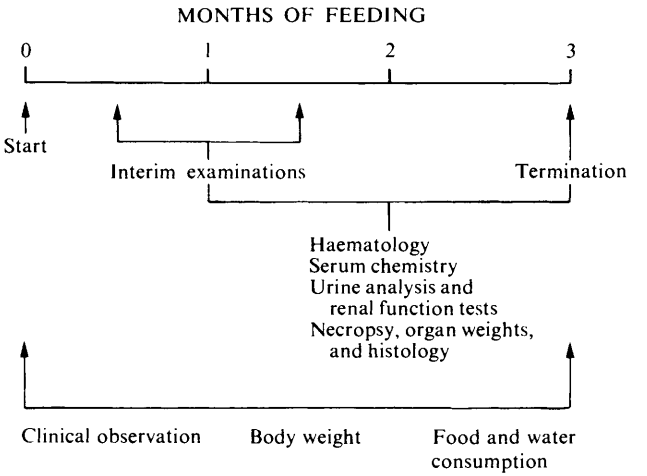

Figure 4 *Progress of a short-term feeding study*

examinations of the animals are performed. On these animals as many parameters as appropriate are measured or observed. These include body weight, food and water consumptions, appearance and behaviour, haematology, serum chemistry, and urine analysis. At the post-mortem examination, the macroscopical appearance of the tissues and the weights of the organs are recorded. Subsequently a histological examination of the tissues is performed. This is an outline of a typical short-term feeding study and obviously the test can be tailored to a particular flavouring material. For example, extra rats may be required in the study so that a special investigation can be performed. An example of this is the measurement of hormone levels (*e.g.* insulin or prolactin) in the blood. Also, extra tissue may be required for electron microscopic examination. Such a 90-day study enables a good assessment in the rat of the effects of a material over a wide dose range, and for a period of time which is approximately equivalent to ten years in the life span of man.

There is a great lack of experimental data on flavouring materials and, naturally, expert committees are grateful for any information. In 1965, Oser and co-workers[45] suggested a 'short-cut' which might appear to generate the necessary data rapidly. They proposed that several materials could be examined in parallel against a single control group of animals. One of each experimental group would receive one of the flavourings at a dose level which was 100 times the

[45] B. L. Oser, S. Carson, and M. Oser, *Food Cosmet. Toxicol.*, 1965, **3**, 563.

maximum, on a weight basis, which would be consumed by man in a day. This study would continue for the usual 90 days. This method is sound as long as no untoward effects are obtained in the control and appropriate experimental groups. However, this is rare in practice. Usually, in a standard 90-day study there are some anomalous results which can be dismissed on examination of the other treatment levels but, of course, this is not possible if the material is administered at only one level. Other disadvantages of this 'short cut' which must be obvious to Oser are: (*a*), when more than three flavours are examined at the same time it is necessary to increase the number of animals in the control group; (*b*), the estimate of a no-untoward-effect level for each flavour in the investigation is poor or impossible. It could be much less or much more than the treatment level employed; (*c*), all the flavourings have to be administered in the same way, *e.g.* in the diet, or drinking water, or by intubation; (*d*), the same solvent has to be employed throughout otherwise the control group is not strictly comparable; (*e*), many flavouring materials in trace amounts can be very pleasant, but in large quantities they are often very unpleasant to the staff and probably to the rats. This was particularly obvious when studying furfuryl mercaptan. While in its stock container the odour which leaks out is a delicious aroma of coffee, but it is quite different when the container is opened. Because of this factor the treatment level which the rat will accept is often self-limiting unless the flavouring is administered by stomach tube; (*f*), finally, such an investigation only gives short-term toxicity information and this type of experiment is completely unacceptable in place of a long-term study when one is looking also for possible carcinogenic effects.

The Oser-type test is an attempt at a 'short cut'. However, far from reducing the standards, these days the tendency is to make the requirements for the assessment of safety more rigorous. This point can be illustrated by the recent FASEB report (1976).[46] Last year a 119-page report of the Select Committee on Flavouring Evaluation Criteria of FASEB (The Federation of American Societies for Experimental Biology) was produced at the request of the FDA. Most of this report is concerned with the ways of determining the priorities for testing flavours. However, it also recommends the type of studies which should be performed and outlines the ways in which they should be carried out. Instead of the 90-day study, the FASEB report suggests that such a study should be carried out in rats from parents which have been exposed to the flavouring material from weaning. Also, some of the animals at the end of the exposure period should have the administration of the test material stopped and the animals should be placed on control diet for a further four weeks to see if any observed effects are reversible. Naturally the introduction of a more rigorous test procedure is to be commended, but it considerably increases the time and cost of performing the study. Also, if the animals do not produce young due to a direct or even an indirect effect of the flavouring then it is obviously impossible

[46] 'Criteria for Evaluation of the Health Aspects of using Flavoring Substances as Food Ingredients'. Federation of American Societies for Experimental Biology, FDA no. 223-75-2002, 1976.

to carry out the main 90-day study. There is something to be said for carrying out separate short-term, reproduction and teratological studies, although this means that the flavouring would not be administered to two generations of animals.

In this country the main, generally accepted, method of ensuring that a flavouring material has no long-term toxic or carcinogenic potential is to administer the material at three treatment levels to groups of at least 50 rats of each sex for not less than two years. Such a study resembles the standard short-term study which has been outlined above. However, these days the tendency is to extend the two year period of treatment to a lifetime. In the FASEB report (1976)[46] 'lifetime' is defined as that time when only 20% of the starting group are still alive. Exposure to the compound is required not just from weaning but from the time of conception. Also, in order to establish negative findings as valid, the report requires that more than half the starting rats should have survived at least 18 months. When indicated, specific tests also may be required on certain compounds for unwanted actions such as the production of cataracts, or the measurement of oxygen consumption as an indication of the basal metabolic rate when a material is suspected of acting on the thyroid gland. In the case of ethyl methylphenyl glycidate, which has been mentioned above as having been reported to cause hind limb paralysis, test animals were placed on a treadmill at intervals throughout the study in an attempt to detect early signs of nerve damage.[44] No such neurological lesions were found in this study. Animal tests should usually be performed on more than one species and it is desirable that one of these species should be non-rodent.

Many flavours are esters and therefore hydrolysis and metabolic studies are particularly relevant. As far as hydrolysis studies are concerned many esters are capable of being broken down rapidly by enzymes, for example in the intestinal wall, or in gastric juice, or in the liver, to yield the parent acid and alcohol. Butterworth and co-workers[47] have considered that by a process of extrapolation it should be possible to reduce significantly the number of studies which need to be carried out on a series of esters, providing that these esters are rapidly hydrolysed *in vivo* to their parent acid and alcohol and also that there are adequate toxicological data on the acid and alcohol. This principle, if feasible, would save a great deal of time and money in the safety testing of flavours. Butterworth and co-workers[47] have been studying a series of allyl esters in order to put this theory to the test. The allyl esters were selected because they are known to cause a very specific kind of damage to the liver, namely periportal necrosis. Detailed studies on allyl alcohol[48] and allyl hexanoate[49] have shown that when these materials are given to groups of rats for 90 days, similar toxic effects are produced by equimolar doses. Thus a comparative study of allyl alcohol and six of its esters was designed.

[47] K. R. Butterworth, F. M. B. Carpanini, I. F. Gaunt, P. Grasso, and A. G. Lloyd, *Brit. J. Pharmacol.*, 1975, **54**, 268P.

[48] F. M. B. Carpanini, I. F. Gaunt, J. Hardy, S. D. Gangolli, and K. R. Butterworth, *Toxicology*, 1978, **9**, in the press.

[49] S. A. Clode, K. R. Butterworth, I. F. Gaunt, P. Grasso, and S. D. Gangolli, *Food Cosmet. Toxicol.*, 1978, **16**, in the press.

Table 11 *Summary of periportal liver lesions produced by allyl alcohol and its esters*

Compound	*Dose* (mg/kg/day)	*Cell enlargement*	*Cell necrosis*	*Fibrosis and bile duct hyperplasia*
Allyl alcohol	5	*	1	1
	25	*	1	8
	60	*	6	5
Allyl acetate	8	*	2	0
	43	*	3	0
	103	*	7	5
Allyl propionate	9	0	1	0
	49	0	1	0
	117	4	5	7
Allyl hexanoate	13	*	1	0
	67	*	4	3
	161	*	4	7
Allyl isobutyrate	11	0	0	0
	55	0	0	0
	132	2	0	0
Allyl isovalerate	12	0	0	0
	61	0	1	0
	149	8	5	1
Allyl 2-ethylhexoate	16	0	0	0
	80	0	0	0
	192	7	3	0

*Effect masked by other damage

Table 11 lists the esters employed in the comparative study. The straight-chain esters are the acetate, propionate, and hexanoate, and the branched-chain ones are isobutyrate, isovalerate, and 2-ethylhexoate. They were administered to rats by oral intubation at three dose levels on an equimolar basis. The top dose was selected as the maximum tolerated dose of allyl alcohol while the lowest dose was chosen as one which was expected to produce no macroscopic effects. After 21 days treatment the animals were autopsied and samples of liver were taken for histopathological investigation. Several effects were observed, such as a reduced rate of gain in body weight, but the results of the microscopic examination of the liver were much more dramatic.

The liver lesions fall into three main categories. It is believed that the lesion develops through three stages. The first signs of damage are cells of increased size. It is thought that these then die, forming areas of cell necrosis in the portal region. Then there is an attempt at repair on the part of the liver, which is indicated by fibrosis and bile duct proliferation. This latter stage is indicative of the most extreme damage. The figures indicate the number of animals out of ten

which showed the lesions. Clearly, the results are similar for allyl alcohol and the straight-chain esters. Also there is a dose–response relationship and in all cases there is a high incidence of the most extreme lesions. In fact, in most cases the damage was so great as to obscure any signs of early damage. The difference in the results from the branched-chain esters and those from the straight-chain ones is immediately obvious. There was virtually no effect seen at the two lower levels of treatment and most of the effects seen at the highest level were confined to the early signs of damage.

The results of these *in vivo* experiments are in agreement with *in vitro* hydrolysis studies which indicate that branched-chain esters are hydrolysed about 100 times more slowly to allyl alcohol than are the straight-chain ones. It would appear that the effects observed are due to allyl alcohol which has been liberated by hydrolysis and that the allyl esters should be assessed as a group rather than as single compounds.

In this experiment the question being asked was: 'Is it possible to reduce the quantity of work needed to assess the toxicity of a series of esters if certain biochemical information is available?' and it would appear to be so, providing that certain requirements are met, namely that: (*a*), the esters are hydrolysed in the alimentary tract and the rate and degree of this hydrolysis are known; (*b*), there are well established toxicological data on the parent alcohol and acids; (*c*), certain limited animal studies are performed, so that it is possible to correlate *in vitro* hydrolysis studies with those conducted *in vivo*.

It is not suggested that this is the way in which all esters should be investigated. These were only preliminary experiments with some allyl esters which were intended to investigate the feasibility of such an approach to the problem of the vast number of flavouring esters which have to be assessed for safety. Other ester series would have to be investigated before a true evaluation of this approach could be made.

Returning to the subject of hydrolysis and metabolic studies, it is important to realize that a metabolite of a substance rather than the parent compound may produce a toxic effect. Most metabolic processes in the body detoxify foreign substances, but this is not always true (*e.g.* acrolein produced by the metabolism of allyl alcohol). Usually what is important is to determine whether the compound is metabolized by a similar route in the experimental animal being employed in the toxicity study to that in man. Briefly, metabolic studies are carried out in order not only to determine the metabolites produced, but also to discover the route followed through the body, any possible storage, and how the material is excreted. The metabolites may be determined using any of the multifarious methods employed in modern chemistry, *e.g.* chromatography and i.r. or u.v. spectrophotometry. Alternatively, labelled materials may be employed both to determine the metabolites and to reveal the route followed through the animal. These days whole body radioautography is becoming a commonly employed method of following the passage of a flavouring material through the animal.

It is worthwhile considering the modern trend of trying to replace experimental

animals by cells. The cells may be isolated animal cells such as human fibroblasts, or HeLa cells, or they may be strains of bacteria. Alterations in genetic material in cells are of increasing interest these days because of their relationship to teratogenicity and possibly to carcinogenicity. In this sense the *in vitro* tests which follow are possible 'short cuts' for animal carcinogenicity studies which are both time-consuming and expensive. The use of cells may be illustrated by the Ames test[50], a DNA repair test,[51] and a cell transformation test.[52]

The Ames test is a bacterial mutation test which is intended to establish the mutagenicity of the test material (*e.g.* a flavouring substance) in mutant strains of *Salmonella typhimurium*. Histidine-dependent bacterial strains are suspended in histidine-free agar to which is added the test or control substance. After two days incubation the colonies are counted. With this system only back-mutated bacteria are able to grow in the absence of histidine so that the number of colonies gives an indication of the mutation rate and an answer is obtained in a matter of days.

The second method,[51] DNA repair, employs either an autoradiographic technique or may involve evaluation by scintillation counting.[53] The relevance of this technique depends on the observation that carcinogens produce damage of the DNA. Repair of this damage can be demonstrated by the uptake of ^{3}H-thymidine by human fibroblasts but only after suppression of scheduled DNA synthesis by growing the cells in an arginine-deficient medium or by blocking the cells chemically by hydroxyurea. Again, using a DNA repair method it is possible to obtain an answer in a matter of a few weeks.

The last method is a cell transformation test. In a living animal, one important characteristic distinguishes cancer cells from normal cells, namely their capacity for uncontrolled proliferation. This property is reflected in the ability of malignant cells to form colonies when inoculated into soft agar; non-malignant cells survive under these conditions but do not multiply to form colonies. Such transformations of normal cells can be brought about by chemical carcinogens *in vitro* and therefore cell transformation *in vitro* has been suggested by Purchase and co-workers[52] as a means of detecting potential chemical carcinogens. Here again we have a simple screening test for carcinogenicity which can be completed in a matter of days.

Of course, there are many limitations to these *in vitro* tests, but in view of the large number of flavouring materials which require safety assessment such tests could permit a rapid screening of a large number of them. In the long run, however, it is still necessary to carry out animal studies in order to assess the risk to man, but *in vitro* tests may provide an early warning of materials which might give rise to concern.

[50] B. N. Ames, J. McCann, and E. Yamasaki, *Mutat. Res.*, 1975, **31**, 347.
[51] R. H. C. San, and H. F. Stich, *Internat. J. Cancer*, 1975, **16**, 284.
[52] I. F. H. Purchase, E. Longstaff, J. Ashby, J. A. Styles, D. Anderson, P. A. Lefevre, and F. R. Westwood, *Nature*, 1976, **264**, 624.
[53] J. E. Troska, and J. D. Yager, *Exp. Cell Res.*, 1974, **88**, 47.

Summary
An adequate assessment of safety of a flavouring material must be based on a series of inter-related investigations which are invariably time-consuming and range from simple chemical tests at one extreme to elaborate biological studies at the other. The object of the present paper is not to suggest definitive methods, but rather to contrast the present accepted methods with the possible procedures of the future. However, such 'short cuts' are unproven and require much further evaluation before it is likely that expert committees will accept them.