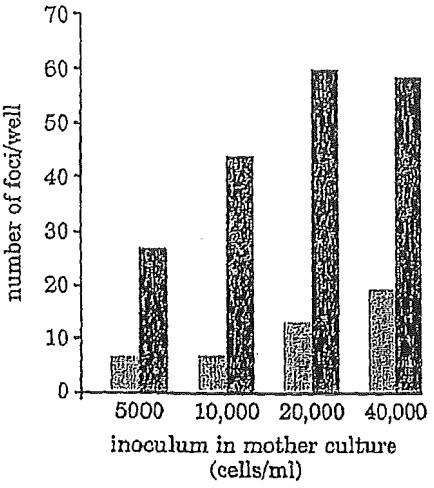
Figure 8: Effect of mother culture inoculum density on spontaneous and TPA-induced transformation frequency



▒ = *DMSO control;* ▒ = *TPA 50ng/ml.*

the results from these three laboratories were judged to be equivocal according to the original criteria (Table 2, column 6).

Issues raised by Study II

Four issues arose from the results obtained in Study II.

1. The selection of test concentrations.
With regard to the selection of test concentrations, the former report (14) recommended that the highest test concentration to be used for an L-type chemical was that showing around 20% cytotoxicity. According to the present results for LCA, it was noteworthy that transformed foci were formed at concentrations showing relatively high growth inhibition: Figure 1b revealed that cytotoxicity at 20μg/ml was about 30%. Thus, a concentration exhibiting more than 30% cytotoxicity was recommended as the highest test concentration.

2. Inter-laboratory variation in response.
In order to address the issue of inter-laboratory variation, various factors influencing culture conditions were examined. It was demonstrated that the use of plates sourced from different manufacturers had little effect (data not shown). When the medium kept in a refrigerator for one month was compared with freshly prepared medium, the numbers of foci were not significantly different. The size of the foci, however, was slightly smaller with the preserved medium (data not shown), suggesting that the use of freshly prepared medium was advisable. Nevertheless, the underlying reason for the inter-laboratory variation remained unclear. According to previous experiments of the management team, the use of M10F medium for mother cultures showed a tendency to induce fewer foci as compared to the use of DF5F medium. Therefore, the original protocol for the collaborative work adopted the use of DF5F medium. After precise checking of each experiment conducted in Study II, it became evident that laboratories 2 and 7 had used M10F medium for the mother cultures. This suggested that this was the cause of the low transformation response in these laboratories. In the subsequent studies, DF5F medium was required for use with mother cultures, in order to reduce this source of variation.

3. Difficulty in scoring foci.
The principal investigators agreed that, while transformed cells were apparent from their basophilic staining and spindle-shape, there were some irregular shaped foci which made scoring difficult. The scoring of such foci as transformed was not consistent between the laboratories. In order to make focus counting much easier, an experiment in which cells were cultured for longer periods was carried out, with 3 wells per experimental condition (Figure 5). In the original protocol (Group A), cells were cultured for 17 days, the last three days of which were with control medium. When culture was continued for another four days in control medium (Group B with medium change, and Group C without medium change), the foci formed grew larger and became easier to score than the foci in Group A, although the number of foci formed increased (Figure 5b; compare the sizes and number of foci in Group A to those in Groups B and C). When 3-week cultures (Groups B and C) were compared, the number of foci was not different between them (Figure 5c). On the basis of these results, it was decided that the protocol should be modified to extend the assay period to three weeks, with no medium change during the last week (Figure 5a, time schedule C).

4. Problems in making judgements.
Some difficulty arose in judging the results for LCA from laboratories 4 and 5 to be equivocal (Figure 3). In these laboratories, a statistically-significant serial increase of focus number was observed. One of the reasons for the judgement was related to the number of foci in the control cultures. When the number of foci in the control cultures was elevated, the third criterion of the original protocol of a three-fold increase above the control seemed too stringent. Incidentally, two-thirds of the laboratories obtained more than five foci in the control cul-

Table 4: Evaluation of the promotion assay of various chemicals, based on two criteria (Study IV, Figure 9 results)

Compound	Lab No.	1: No. of concentrations with significant effect	4: No. of concentrations showing a 2-fold increase	Judgement based on 1 + 4
Progesterone	5	4	3	+
	7	6	5	+
	9	5	5	+
	13	4	5	+
Diethylstiboestrol	5	0	0	-
	7	0	0	-
	9	0	0	-
	13	0	0	-
Anthralin	1	1	0	±
	6	0	0	-
	8	1	0	±
	12	0	0	-
Insulin	1	3	2	+
	6	4	3	+
	8	5	4	+
	12	4	3	+
Catechol	3	3	3	+
	4	0	0	-
	10	2	4	+
	14	4	3	+
Sodium saccharin	3	3	3	+
	4	3	3	+
	10	4	6	+
	14	5	5	+

ture. In the screening of chemicals, attention should be paid to eliminating false negatives in the judgement. From this consideration, the three-fold increase requirement was reduced to a two-fold increase requirement (giving a fourth criterion). Of course, it remains essential that every effort be made to maintain an appropriate number of foci in the control cultures (see *Issues raised in Study III*, below).

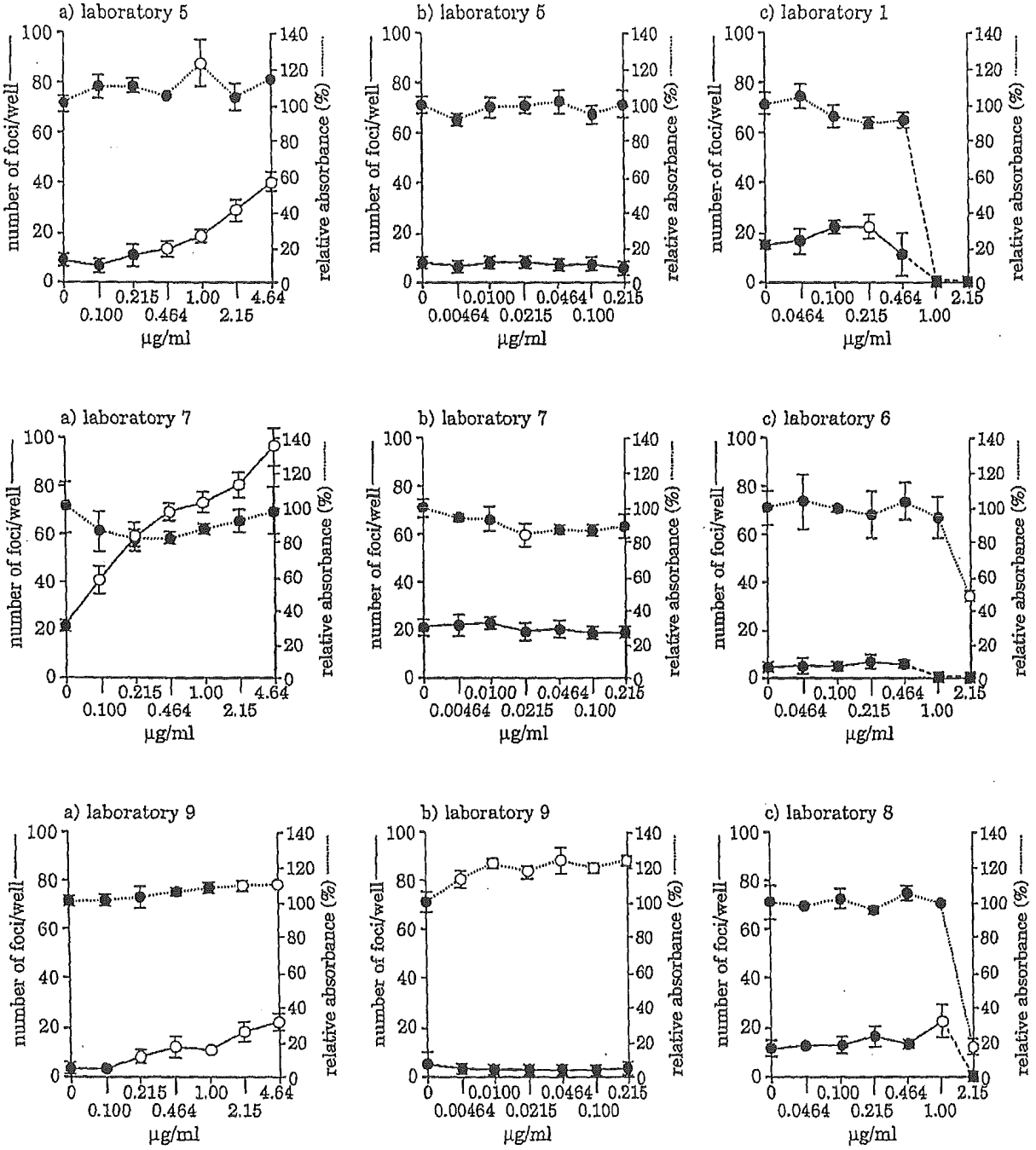
In addition, the second criterion of a significant increase at consecutive concentrations posed some difficulty. For example, although the patterns of responses to LCA in laboratories 7 and 10 were similar, the former result was judged to be positive, while the latter was judged to be equivocal (Figure 3). Laboratory 10 observed a clearly significant increase in transformed foci, but only at one concentration. Thus, while strict criteria are important to avoid including incidental increases in transformed foci, it is sometimes difficult in practice to determine appropriate test concentrations within a narrow concentration range for some types of chemicals. Again, in order to eliminate false negatives, the second criterion was not included in the revised criteria.

Thus, although the criteria were improved, a repeat experiment has to be considered when it is not clear what the judgement should be. As a result of these deliberations, the judgements in Table 2, column 6, were revised. Column 7 represents judgements made according to the revised criteria. All the laboratories obtained a positive result for LCA. For the same reason, a column 7 was also added to Table 1 (relating to TPA), indicating the effects of basing the judgement on the first and fourth criteria, although, in this case, the result was to confirm the judgements shown in column 6.

Bhas promotion assay with various chemicals (Study III)

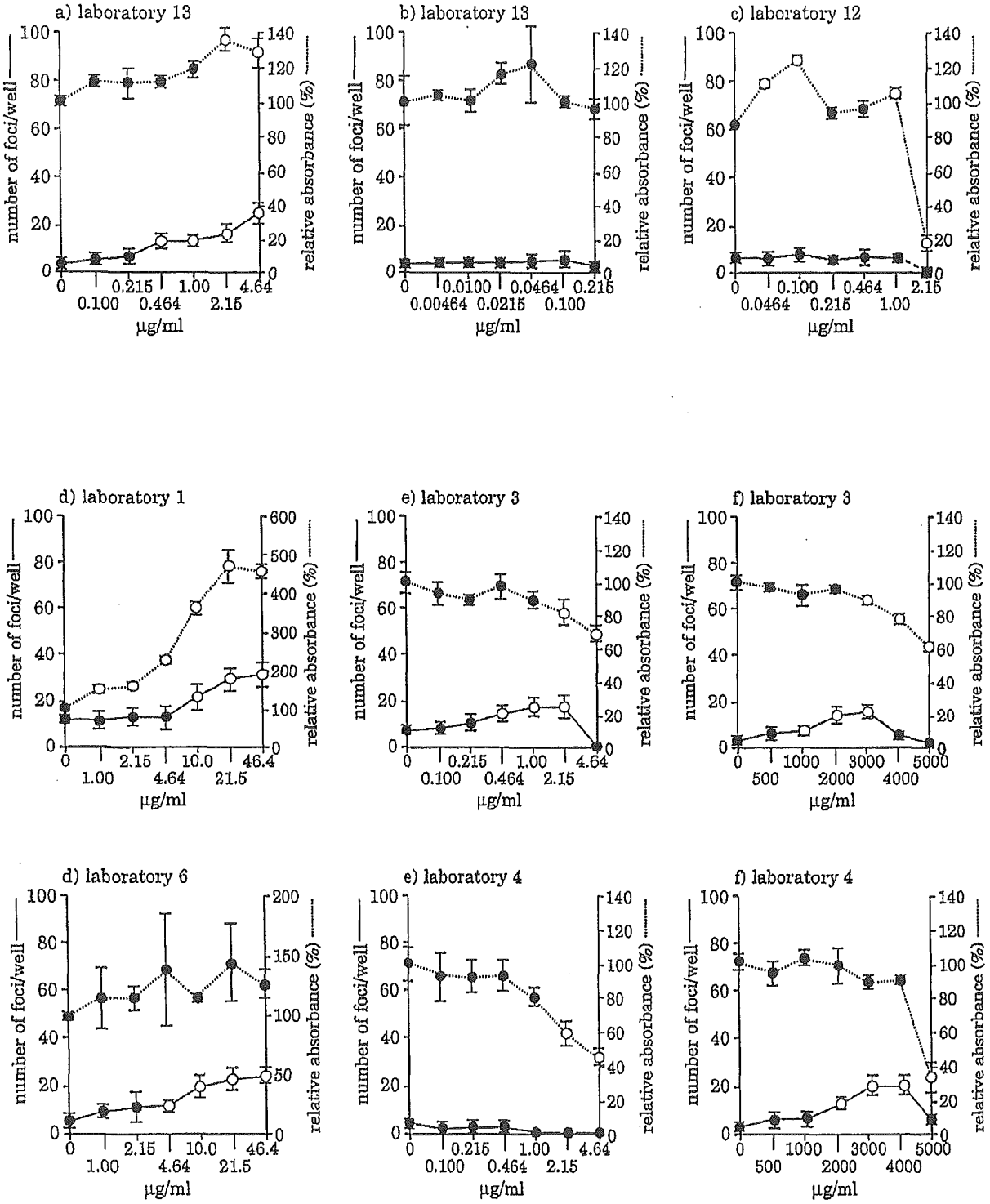
Following the modified promotion assay protocol and the newly-adopted judgement criteria, six chemicals were examined: mezerein, 4α-phorbol, PDD, 17β-oestradiol, okadaic acid, and dexametha-

Figure 9: Transformation frequencies and effects on cell growth with various chemicals (Study IV)



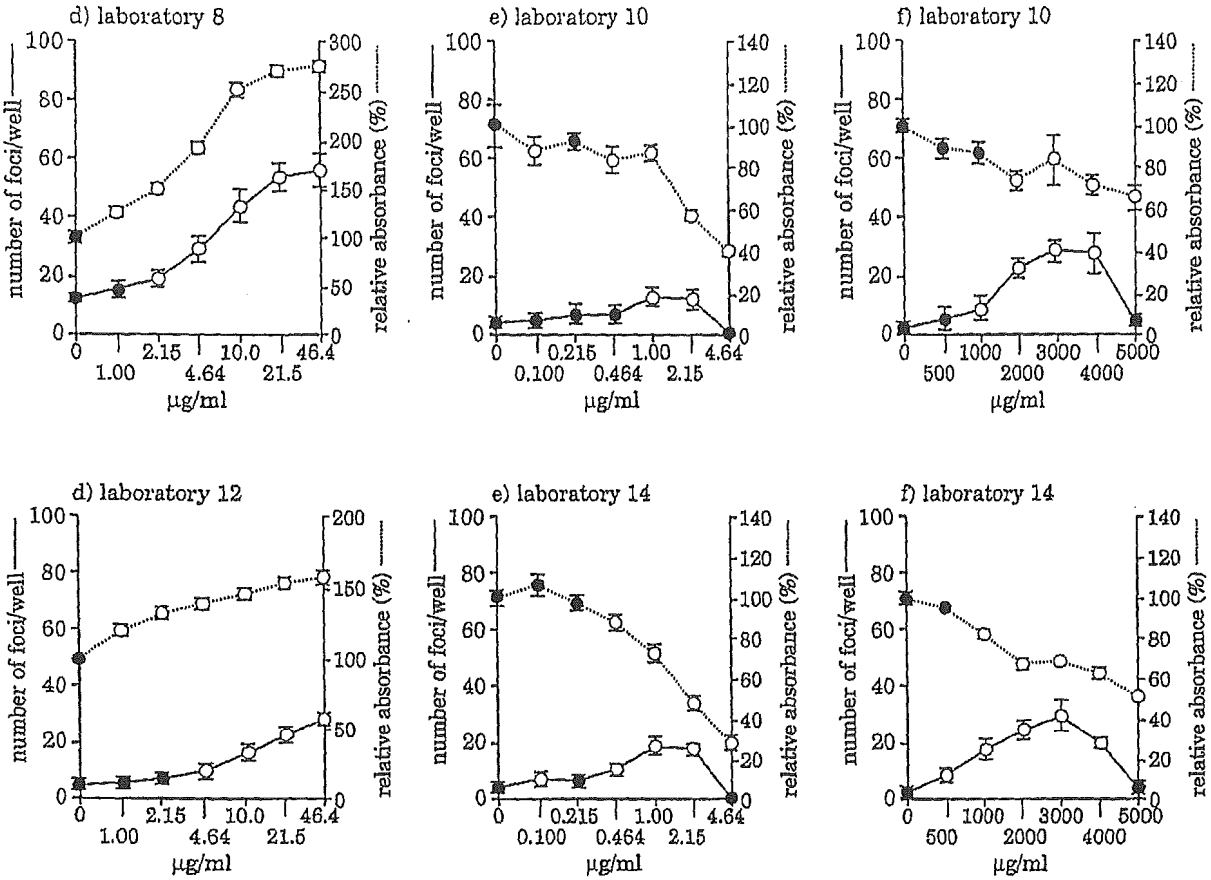
a) progesterone; b) DES; c) anthralin. ○ = *significant increase in focus number in transformation assay, and significant point in cell growth assay compared to control;* ■ = *point of cell death because of toxicity.*

Figure 9: continued



a) progesterone; b) DES; c) anthralin; d) insulin; e) catechol; f) sodium saccharin. ○ = *significant increase in focus number in transformation assay, and significant point in cell growth assay compared to control;* ■ = *point of cell death because of toxicity.*

Figure 9: continued



d) insulin; e) catechol; f) sodium saccharin. ○ = *significant increase in focus number in transformation assay, and significant point in cell growth assay compared to control;* ■ = *point of cell death because of toxicity.*

sone. Twelve laboratories were divided into three groups of four by using a randomised block design based on the focus formation response in Study II. Each group was responsible for testing two chemicals under blind conditions. Again, a randomised block design was used for the assignment of the two chemicals, based on the previous transformation data (14).

The results are shown in Figure 6, and the judgements made in the light of the results are summarised in Table 3. All four laboratories obtained positive results with mezerein and PDD. The results of growth assays on both chemicals indicated T-type chemicals in most of the laboratories. 4α-phorbol and 17β-oestradiol were judged as negative in all four laboratories. In the case of okadaic acid, three laboratories obtained positive results and one obtained negative results. The cell growth assays on okadaic acid indicated it to be an L-type chemical. The results for dexamethasone were divided: two laboratories obtained negative results, while the other two obtained equivocal or positive results. This mixed result is discussed below.

Issues raised in Study III

Two main issues arose from Study III: 1) there was a large inter-laboratory variability of focus formation in the negative control; and 2) there were inconsistent results among the laboratories.

With regard to the first issue, the number of foci in the negative control was over 20 per well in nine of 24 experiments, whereas nine experiments gave less than 10 foci per well. It was considered that this might be related to the use of DF5F medium for the mother cultures. A clue was found in the following experiment. Cells at different cell numbers

(5–40 × 10^3 cells/ml) were prepared for mother cell culture. The state of cell growth after three days is shown in Figure 7. The growth of cells inoculated at a lower cell density was sparse (Figure 7a), but increased with inocula with higher cell numbers (Figure 7b to 7d). Cells harvested from these mother cultures were assayed for TPA promotion, with three wells per experimental condition (Figure 8). The number of foci in negative controls was lowest, when 5–10 × 10^3 cells/ml had been used in the mother culture inocula, and increased 2-fold to 3-fold for cells derived from higher density mother cultures. Similarly, the number of foci in cultures treated with TPA increased, dependent on the cell density of mother cultures, but plateaued in assays involving cells derived from confluent mother cultures. Thus, it was apparent that focus formation was dependent on cell density within mother cultures, at the time of cell harvest and establishment of the assay. Though the reason for this remains unclear, it is evident that attention should be paid to the cell numbers plated in mother cultures. For future experiments, a decision was made to use inocula of 1 × 10^4 cells/ml (a total of 1 × 10^5 cells per 9cm dish) and three days of culture, before preparing cells for use in assays.

The second issue was the inconsistency of results among some of the laboratories. In the case of okadaic acid, a known tumour promoter, three laboratories gave positive results, but one laboratory obtained a negative result. Making judgements about toxic chemicals such as okadaic acid and LCA can be difficult, and requires special attention. Dexamethasone was judged as negative in two laboratories, but laboratory 1 obtained an equivocal result and laboratory 5 obtained a positive result. In laboratory 1, a statistically-significant serial increase in the number of transformed foci was observed, but it did not satisfy the revised criterion of a 2-fold increase, probably due to the high number of foci in the negative control culture. In laboratory 5, there were significant, 2-fold increases in transformed foci at two test concentrations, but they were not consecutive. In this case, the number of foci in the negative control culture was low. These results suggest a need for repeat experiments, when equivocal or biologically suspicious results are obtained.

Further Bhas promotion assay on other chemicals (Study IV)

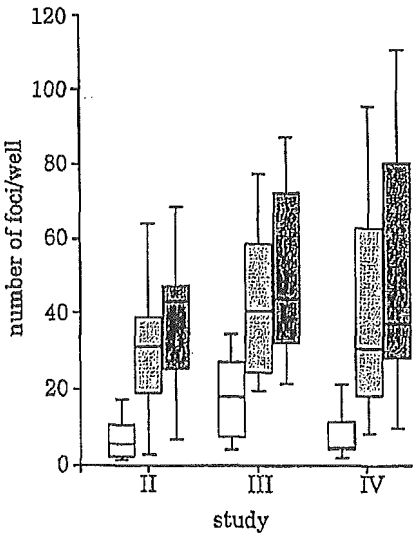
Study IV was conducted following the decision to use a specific cell density for mother cultures (see above). Again, twelve laboratories were divided into three groups, and each group examined two chemicals. The test chemicals were progesterone, DES, anthralin, insulin, catechol and sodium saccharin. The results are shown in Figure 9. This time, only two experiments obtained more than 20 transformed foci per well in the negative control. The number of experiments showing 10 to 20 foci per well in the control was four, and there were less than 10 foci in 18 experiments.

Table 4 summarises the judgements based on the results. All four laboratories testing progesterone, insulin and sodium saccharin judged these chemicals to be positive. Cell growth assays identified insulin as a T-type chemical, and progesterone and sodium saccharin as M-type chemicals. Negative results were obtained by all four laboratories for DES. In the case of anthralin, two laboratories obtained equivocal results, and the other two judged the results to be negative. Catechol was judged to be positive in three laboratories and negative in one laboratory.

Issues raised in Study IV, and overall discussion

The issue raised in Study III concerning the high number of foci in the controls was markedly improved in Study IV. Eighteen out of 24 experiments gave less than 10 transformed foci per well in the negative control culture, compared to only 9 out of 24 in Study III. However, the fact that there were two experiments with more than 20 foci per well in the controls remains cause for concern. Clearly, vig-

Figure 10: Box-whisker plot of transformation frequencies with TPA in Studies II, III and IV



□ = *DMSO control;* ▨ = *TPA 20ng/ml;* ▩ = *TPA 50ng/ml.*

ilance is needed with the cell maintenance conditions.

In the experiments in Studies III and IV, treatment with 20ng/ml and 50ng/ml TPA was routinely performed as a positive control. The negative control and the TPA-treatment data were combined and are shown as box-whisker plots in Figure 10, together with Study II data. Inter-laboratory variation in results for the negative controls was considerable in Study III, but much reduced in Study IV. The responses to TPA were relatively constant and reproducible among the three studies.

With regard to the second issue raised in relation to Study III, two chemicals also produced inconsistent results in Study IV (anthralin and catechol). Nonetheless, overall, eight out of 12 chemicals in Studies III and IV gave consistent results in all the four laboratories concerned. Moreover, for two of the other four chemicals, only one laboratory out of four showed inconsistent results. Therefore, the rate of consistency appears to be quite high. From the viewpoint of practical chemical screening, however, the presence of any inconsistency is unfavourable. The fact that both positive and negative results were obtained with okadaic acid, dexamethasone and catechol, poses a problem for the reliability of this assay. This assay requires further improvement and technical refinement.

17β-oestradiol (15) and DES (16) have been reported to be carcinogenic in animal studies, but there are no published studies on the tumour-promoting activity of these oestrogenic chemicals. The *in vitro* assay results on these chemicals in this study were negative.

It was concluded that three different types of chemicals showed positive promoting activity in the Bhas assay. T-type chemicals clearly enhanced cell growth, usually by more than 150% compared to growth in untreated control cultures. The T-type chemicals were TPA, mezerein, PDD and insulin, and they induced transformed foci at concentrations exhibiting growth enhancement. LCA and okadaic acid were classed as L-type chemicals, which were toxic at a narrow range of concentrations and induced transformed foci at concentrations showing slight to 70% growth inhibition. In the case of M-type chemicals, progesterone induced transformed foci at test concentrations with little or no growth inhibition, whereas catechol and sodium saccharin showed toxicity over a wide range of concentrations and induced foci at concentrations with slight to little growth inhibition. Thus, M-type chemicals are distinctly different from L-type chemicals. The further accumulation of data on many other chemicals is necessary to consolidate this categorisation, but this approach may provide clues of value in elucidating mechanisms of tumour promotion.

Recently, Asada *et al.* have found that Bhas 42 cells can also detect the initiating activities of chemicals (17). It is obviously advantageous that this sensitive Bhas transformation assay can detect both initiating and promoting activities. A further inter-laboratory collaborative study on use of the Bhas assay for detecting both the initiating and promoting activities of chemicals is now being planned. Further developments will also include mechanistic studies, as well as the technical transfer of the methodology to other laboratories, and a wide range of validation studies.

Acknowledgement

This study was supported by a Grant-in-Aid from the Japan Chemical Industry Association.

Received 4.3.05; received in final form 24.6.05; accepted for publication 20.7.05.

References

1. Sugimura, T. (1992). Multistep carcinogenesis: a 1992 perspective. *Science, New York* **259**, 603–607.
2. OECD (1997). *Guidelines for Testing of Chemicals. Section 4. Health Effects.* Paris, France: Organization for Economic Co-operation and Development. Website http://www.oecd.org/document/55/0,2340,en_2649_34377_2349687_1_1_1_1,00.html
3. Kakunaga, T. (1973). A quantitative system for assay of malignant transformation by chemical carcinogens using a clone derived from BALB/3T3. *International Journal of Cancer* **12**, 463–473.
4. Reznikoff, C.A., Bertram, J.S., Brankow, D.W. & Heidelberger, C. (1973). Quantitative and qualitative studies of chemical transformation of cloned C3H mouse embryo cells sensitive to postconfluence inhibition of cell division. *Cancer Research* **33**, 3239–3249.
5. Hallowell, A., Tu, W., Pallotta, S., Sivak, A., Lubet, R.A., Curren, R.D., Avery, M.D., Jones, C., Sedita, B.A. & Huberman, E. (1986). An interlaboratory comparison of transformation in Syrian hamster embryo cells with model and coded chemicals. *Environmental Mutagenesis* **8**, 77–98.
6. Tsuchiya, T. & Umeda, M. (1995). Improvement in the efficiency of the *in vitro* transformation assay method using BALB/3T3 A31-1-1 cells. *Carcinogenesis* **16**, 1887–1894.
7. Yotti, L.P., Chang, C.C. & Trosko, J.E. (1979). Elimination of metabolic cooperation in Chinese hamster cells by a tumor promoter. *Science, New York* **206**, 1089–1091.
8. Murray, A.W. & Fitzgerald, D.J. (1979). Tumor promoters inhibit metabolic cooperation in co-cultures of epidermal and 3T3 cells. *Biochemical and Biophysical Research Communications* **91**, 395–401.
9. Umeda, M., Noda, K. & Ono, T. (1980). Inhibition of metabolic cooperation in Chinese hamster cells by various chemicals including tumor promoters. *Japanese Journal of Cancer Research* (Gann) **71**, 614–620.
10. Rovera, G., Santoli, D. & Damsky, C. (1979). Human promyelocytic leukemia cells in culture differentiate into macrophage-like cells when treated with a phor-

bol diester. *Proceedings of the National Academy of Sciences of the USA* **76**, 2779–2783.
11. Ito, Y., Yanase, S., Fujita, J., Harayama, T., Takashima, M. & Imanaka, H. (1981). A short-term *in vitro* assay for promoter substances using human lymphoblastoid cells latently infected with Epstein-Barr virus. *Cancer Letters* **13**, 29–37.
12. Ohmori, K., Miyazaki, K. & Umeda, M. (1998). Detection of tumor promoters by early antigen expression of EB virus in Raji cells using a fluorescence microplate-reader. *Cancer Letters* **132**, 51–59.
13. Busser, M.T. & Lutz, W.K. (1987). Stimulation of DNA synthesis in rat and mouse liver by various tumor promoters. *Carcinogenesis* **81**, 1433–1437.
14. Ohmori, K., Sasaki, K., Asada, S., Tanaka, N. & Umeda, M. (2004). An assay method for the prediction of tumor promoting potential of chemicals by the use of Bhas 42 cells. *Mutation Research* **557**, 191–202.
15. Nagasawa, H., Mori, T. & Nakajima, Y. (1980). Long-term effects of progesterone or diethylstilbestrol with or without estrogen after maturity on mammary tumorigenesis in mice. *European Journal of Cancer* **16**, 1583–1589.
16. Huseby, R.A. (1980). Demonstration of a direct carcinogenic effect of estradiol on Leydig cells of the mouse. *Cancer Research* **40**, 1006–1013.
17. Asada, S., Sasaki, K., Tanaka, N., Takeda, K., Hayashi, M. & Umeda, M. (2005). Detection of initiating as well as promoting activity of chemicals by a novel cell transformation assay using v-Ha-ras-transfected BALB/c3T3 cells (Bhas 42 cells). *Mutation Research* **588**, 7–21.

Environ. Mutagen Res., 27: 57-60 (2005)

統計家からの疑問：in vitro 試験のデータから定量的結論を導くのは邪道か？

吉村　功*

東京理科大学工学部　〒162-8601 東京都新宿区神楽坂1-3

Why don't we use in vitro assays for quantitative risk assessment

Isao Yoshimura

Faculty of Engineering, Tokyo University of Science, 1-3, Kagurazaka, Shinjuku-ku, Tokyo 162-8601, Japan

Summary

This paper discusses, from statistical viewpoint, about the methodology to examine whether substances have threshold on tumorgenecity or genotoxicity based on experiment data. When we try to discriminate models with threshold from those without threshold through in vivo experiment, the necessary sample size n will be too huge to be practical. In vitro assays seem to be more promising. Current methodologies utilized in in vitro assays, however, are not suitable for our purpose because they are involved in the determination of the positivity/negativity of test chemicals, while the methodologies we require are those for estimating dose-response relationships. Collaborative studies among toxicologists and statisticians are desirable to establish reasonable methodologies.

Keywords: dose-response relationship, genotoxicity in vitro assay, statistical analysis, threshold

緒　　言

著者は統計家である．著者が毒性試験のデータ解析の相談を受けるようになったのは約25年前(吉村, 1987)である．毒性試験の共同研究等に深く関わるようになったのは，約15年前の小核試験の統計的データ解析(Hayashi et al., 1989)からである．その後，多種多様な in vivo 試験と in vitro 試験のデータ解析を手がけてきた(参考文献参照)．今回述べるのはこの経験の中で感じた，in vitro 試験のデータ解析についての疑問と，それに対する著者の意見である．

1．閾値の有無の実験的確認という問題

閾値があればADI(admissible daily intake)を定め易くなることや，閾値があるという認識を強調するのはリスクを過小評価するためという批判があることから，閾値があるかどうかが関係者の間で大きな研究課題となっている．この「発がん性と遺伝毒性の閾値・リスクアセスメントにおける問題点」というシンポジウムもこの課題の研究の一環であろう．

著者がこの問題について最も重要と考えているのは，高用量で毒性(発癌性や遺伝毒性)が認められている物質(放射線も含める)の低用量閾値の有無が，実験データで確認できるのか，もしできるとすればそのための方法はどのようなものかである．

放射線に関しては，広島の疫学調査がヒトについてのデータを出しているが，これは方法論として例外である．一般には，in vivo/in vitro 毒性試験で閾値の有無を確か

* E-mail: isao@ms.kagu.tus.ac.jp

受付: 2005年5月7日　受理: 2005年5月9日

本稿は第33回日本環境変異原学会，第18回日本動物実験代替法学会合同学術会議，JEMS & JSAAE合同シンポジウム1「発がん性と遺伝毒性の閾値—リスクアセスメントにおける問題点—」で発表された．
This paper was presented to the JEMS & JSAAE combination symposium 1 "Threshold of carcinogenicity and genotoxicity — issues for risk assessment —" at the 33rd JEMS annual meeting and the 18th JSAAE annual meeting, 2004.

められるかどうか，考えるべきである．

2．In vivo データでの確認の困難性

統計的視点で考えた場合，閾値の存在を検討するin vivo試験のやり方は用量群実験である．すなわち，0用量(溶媒対照など)を含めた複数水準(a水準)の用量を設定し，各用量に対してn匹の動物からなる実験群を用意し，各用量の被験物質を曝露させ，問題となっている反応の出現率を観測し，各用量群における出現率から，母集団での用量反応関係を推定することである．

この用量群実験で閾値を確認しようとするなら，(M1)ある用量以下では，母出現率(実験の背景にある動物集団での出現率)が0用量での母出現率と同じであるとするモデル(ホッケースティック回帰モデル)が，データに最も良く適合することを示す，(M2)ある用量以下では，母出現率が0用量より小さくなることを示す，(M3)ある用量以下では，母出現率がある値Δ(virtually safe level)以下であることを示す，ことであろう．これ以外の方法論は現存しないと著者は考えている．

この3つの方法の内，(M2)は，例えば低用量だとがんが抑制される，といった非常に特殊な被験物質以外には適用できないので，例外とすべきであろう．

そこで対象とできるのは，(M1)，(M3)ということになる．しかし，「0用量群と有意差がなければその群での母出現率は0用量と同じと考えてよい」という論理は，統計学の世界で否定されている．したがって，(M1)で積極的に，閾値の存在を証明できた，というわけにいかない．ぎりぎりで，「モデル選択の手法を適用したところ，閾値の存在が示唆された」というところまでである．

となると考えられるのは，(M3)の論理であるが，母出現率をppmのオーダーで考えよう，という現今の議論レベルで，この方法論を適用することは，必要なn数が大きくなりすぎて，一般的方法論として語りにくい．特に放射線のように低用量，低反応が大きな争点になっている問題では，きわめて困難と感じられる．方法論として期待できるとは言い難い．

このような思考過程の後で，著者が現在持っている見通しは，「in vitro試験で被験物質の用量反応関係を評価・推定しておいて，そこで引き起こされる反応が，生体内の修復・防御機構で抑えられることが示せれば，生体としては閾値が出現する」という論理である．このような論理は荒唐無稽だろうか．

3．In vitro 試験のデータ解析法の役割

各種のIn vitro 試験の共同研究での経験によると，実験を行う人(実験家)の多くは，統計的データ解析に，2つの視点を，動かせないものとして示してくる．

その1つは，統計的データ解析は参考のために行うのであって，最終的判断は生物学的に行う，というものである．

もう1つは，これらの試験に求めるのは，被験物質が「陽性positive」か，「陰性negative」か，あるいは「擬陽性equivocal」か，という判断であって，用量反応関係の推定，閾値や無影響量の推定などの量的判断ではない，というものである．

前者について，著者があらゆる機会に強調することは，次のようなことである．

生物学的判断は主観・経験に基づく関係で，個人差が大変大きい．これに頼りすぎることは不毛な議論を導く．統計手法というと，無条件に○○検定といったものを考える実験家が多いが，著者が提示しようとするものはそのようなものではない．熟達した実験家・生物学者の判断を手続きとして客観化して，論理的にその妥当性を吟味したものである．客観的手続きで判定を下すことは，それを無条件に受け入れるべきだということではなく，これを通して主観的・経験的判断の弱点を補い，個人差による判断の違いを減らすことである．これを生物学的な仕組みについての知見と照らし合わせることが，科学的判断のレベルをあげることになるのではなかろうか．

実際，このような視点で手法を開発しようとすると，ときに，大変興味深い現象に出会う．例えばMLA(mouse lymphoma assay)の場合，統計モデルを設定してモンテカルロシミュレーションでデータ解析法の性能を比較したときと，実際の実験データでの熟達した実験家の判定を取り入れてデータ解析法の性能を比較したときで，データ解析法の性能評価の結果が食い違うのである．近いうちに，このデータを提示する予定であるが，この食い違いの合理的な説明について，実験家と統計家の間での議論が必要と著者は考えている．

4．統計的有意性と生物学的有意性

前項で述べた方針で適切な感度・特異度を保持する統計手法を考案したとする．その手法を検定という視点で見ると，有意水準つまり偽陽性確率を非常に小さくしたものが得られる．

すなわち，統計学的モデルとしては，誤差的でない単調増加性が認められるデータに対して，実験家は「勾配が小さいから陰性」「最大反応値が小さいから陰性」と判断している場合が多いのである．

これは何故だろうか．これは，著者がずっと抱き続けている疑問である．これに対する，著者が十分納得できる明快な回答は，今のところ得られていない．

著者の解釈は次のようなものである．

実験家がバリデーション研究で実験に用いる被験物質は，マウス・ラット・ヒトなどで，毒性がある程度わかっているものが多い．

そのような被験物質でin vitro試験を行ったときには，動物等で毒性が認められない被験物質で，「この程度の

勾配で用量と共に反応が増加する」,「この程度は溶媒より反応の最大値が大きくなる」,「この程度の用量反応関係は，ある実験では出るが，他に実験を繰り返すと現れなくなる」ことが多い．

実験家はこのような経験・知見を基にして，in vitro試験データの結果を評価しているのではないだろうか．そうであれば，実験家は，in vitro試験の結果をin vivo試験の結果に依拠して評価していることになる．それは，動物生体での防御機構・修復機構・閾値構造を考慮に入れてin vitro試験のデータの解析をしていることである．実験家の判断にそういう側面があり，それを生物学的有意性と表現しているのではないだろうか．

5. In vitro試験での定量的判断の意義

現実の実験が用量群実験であるにもかかわらずin vitro試験では，陽性・陰性・擬陽性，の判断のみが求められるのはなぜだろうか．いろいろ実験家に尋ねているが，まだ納得できる回答に出会っていない．

著者の疑問に対する比較的多い回答は，in vitro試験の結果の再現性は，定量的結論を導くだけの性能を持っていない，というものである．本当だろうか？　もし本当だとすると，では，in vivo試験では再現性が十分か，という疑問が生じる．それは程度の問題にすぎないのではなかろうか．

それにもかかわらず，in vivo試験のデータ解析では，用量反応関係を全体として評価している．そして，閾値の存在，無作用量の推定，LD50で毒性の強さを表示している．であれば，in vitro試験でも，ある範囲で定量的結論を導く場合があって良いのではないかと著者は考えている．

定量的結論を導くことが有用な場合は2つ考えられる．1つは，代替法としてin vitro試験を利用する場合，他の1つは本シンポジウムの主題となっている閾値の有無の評価の場合である．

前者については，例えば眼刺激性試験代替法のガイドライン作成の際に議論をしたことがある．これをさらに他の試験の代替法としての性能評価に使う，という利用になる．

後者については，たとえば，in vitro試験のデータで被験物質の用量反応関係を推定して，閾値の存在の有無を議論する．閾値が存在しないという結果の場合には，in vitro試験で認められた毒性反応の生体における発現に，防御機構・修復機構が存在するかどうかをin vivo試験で検討することになる．閾値があれば，それはそれで一つの証拠になる．この場合のin vitro試験の強みは，nを大きくできることが多いことである．

結　語

毒性発現率をppmというオーダーで評価する場合に，in vivo試験で閾値を特定することは，統計的にほとんど不可能である．可能性があるとすると，in vitro試験あるいはそれに準じる(小核試験のような)in vivo試験で評価することであろう．

もし，この判断を認めるならば，そのときの隘路は，in vitro試験において用量反応関係を定量的に評価する習慣と方法論が確立していないことである．

閾値の有無を実験を通して評価するには，in vitro試験を，陽性・陰性の議論の枠に留めないで，用量反応関係の定量的評価に進めなければならない．その可能性を探るには，試験家と統計家の共同研究が必要である．今後に期待したい．

倫理規定

本論文では，人権，個人情報保護法あるいは動物愛護法等の規定にふれる問題を取り扱っていない．

参考文献

(個別には引用していないが著者が関係した毒性試験の研究論文と著書)

Hayashi, M., M. Ishidate, T. Sofuni and I. Yoshimura (1989) A procedure for data analysis of the rodent micronucleus test involving a historical control, Environmental and Molecular Mutagenesis, 13, 347-356.

Hayashi, M., S. Hashimoto, Y. Sakamoto, C. Hamada, T. Sofuni and I. Yoshimura (1994) Statistical analysis of data in mutagenicity assays: in the case of rodent micronucleus assay, Environmental Health Perspectives, 102, Suppl.1, 49-52.

Hamada, C., M. Nomura, K. Matsumoto, I. Abe, K. Yoshino and I. Yoshimura (1998) Tree type algorithm for statistical analysis in chronic toxicity studies, Jour. Toxicological Sciences, 23, 173-181.

Hamada, C, et al. (1997) Detection of an outlier and evaluation of its influence in chronic toxicity studies, Drug Information Journal, 32, 201-212.

Hamada, C., et al. (1997) A study on the consistency between statistical evaluation and toxicological judgment. Drug Information Journal, 31, 323-326.

Hauschke, D., et al. (1997) Recommendation for biostatistics of mutagenicity studies, Drug Information Journal, 31, 413-421.

Hasegawa, R., K. Imaida, N. Ito, T. Shirai and I. Yoshimura (1996) Analysis of synergism in carcinogenesis based on preneoplastic of induction in the rat liver, Japanese Journal of Cancer Research, 87, 1125-1133.

Hirabayashi, Y., et al. (2003) Evaluation of nonthreshold leukemogenic response to methyl nitrosourea in p53-deficient C3H/He mice, Toxicology and Applied Pharmacology, 190, 251-261.

Ilse-Adler, et al. (1998) Recommendations for statistical designs of in vivo mutagenicity tests with regard to subsequent statistical analysis. Mutation Research, 417, 19-30.

Kojima, H., S. Hanamura, A. Miyamoto, H. Sato and I. Yoshimura (1995) Evaluation of seven alternative assays on the main ingredi-

ents in cosmetics as predictors of Draize eye irritation scores, Toxicology in Vitro, 9, 333-340.

Moore, M.M., et al. (2003) Mouse lymphoma thymidine kinase gene mutation assay, International Workshop on Genotoxicity Tests Workgroup Report, Plymouth, UK 2002, Mutation Research, 540, 127-140.

Matsunaga, N., J. Kanno and I. Yoshimura (2002) A statistical method for judging synergism: application to an endocrine disruptor animal experiment, Environmetrics, 14, 213-222.

日本トキシコロジー学会(編)(2003)トキシコロジー用語事典，じほう，東京.

西山　智，吉村　功(2004)．複合最大対比法の提案とその毒性試験データ解析への応用，計量生物学，25，1-18.

Nishiyama, H., T. Tsuchiya, T. Omori, M. Umeda and I. Yoshimura (2002) A composite statistical procedure for evaluating genotoxicity using cell transformation assay data, Environmetrics, 14, 183-192.

大野忠夫(編)(1994)動物実験代替法マニュアル，共立出版，東京.

Omori, T., et al. (1998) Validation study on five cytotoxicity assays, by JSAAE-II, Statistical analysis. Alternatives to Animal Testing and Experimentation, 5, 39-58.

Omori, T., M. Honma, M. Hayashi, Y. Honda and I. Yoshimura (2002) A new statistical method for evaluation of L5178Y tk $^{+/-}$ mammalian cell mutation data using microwell method, Mutation Research, 517, 199-208.

Sonoda, I., et al. (2002) A prevalidation study for three-dimensional cultured human skin models as alternatives to skin irritation testing, Alternatives to Animal Testing and Experimentation, 9, 91-106.

鞆城　豊，他(1996)皮膚刺激性試験における動物数削減—2段階判定法，日本香粧品科学会誌，20，166-172.

Tsuchiya, T., et al. (1999) An inter-laboratory validation study of the improved transformation assay employing Balb/c 3T3 cells: Results of a collaborative study on the two-stage cell transformation assay by the non-genotoxic carcinogen study group, ATLA, 27, 685-702.

Yoshimura, I. and K. Matsumoto (1994) A note on the use of historical controls, Environmental Health Perspectives, 102, Suppl.1, 19-23.

吉村　功(編)(1987)毒性・薬効データの統計解析，サイエンティスト社，東京.

吉村　功，大橋靖雄(編)(1992)毒性試験データの統計解析，地人書館，東京.

渡辺佳津子，坂本京子，佐々木俊明，吉村　功(1996)細菌を用いる復帰変異試験のデータ解析における判定方法の検討：ネズミチフス菌TA102，TA3638，および大腸菌WP2/pKM101，WP2ivrA/pKM101株を用いた共同研究のデータにもとづく評価，Environ. Mutagen Res., 18, 137-160.

Wakana, A., C. Hamada and I. Yoshimura (1997) A performance comparison of maximum contrast methods to detect dose dependency, Drug Information Journal, 31, 423-432.

A Simulation Study for a Linear Measurement Error Model When Error Variances Vary Between Measurements

Yasutaka CHIBA, Yutaka MATSUYAMA, Tosiya SATO, and Isao YOSHIMURA

When predicting scores in the Draize eye irritation test based on measurements of in vitro alternative tests, we are often faced with estimating parameters in a linear measurement error model with heterogeneous error variances. This article proposes a new statistical method for parameter estimation to address this issue. The proposed method is an extension of an earlier proposal that applied a linear measurement error model with homogeneous error variances, to cases with heterogeneous error variances. A simulation study to examine the performance of the proposed method was conducted in a framework that was adaptable to the data, which was obtained in a validation study of alternative methods to animal experiments conducted in Japan. The proposed method reduced the biases of estimates in comparison with an ordinary regression analysis method and three other methods under the assumption of homogeneous error variances. Although the proposed method did not fit the real data well, the resulting prediction formula was far better than those obtained by other methods.

Key Words: Alternative method to animal experiments; Chorioallantoic membrane assay; Draize eye irritation test; Estimating function; Functional relationship model.

1. INTRODUCTION

In the development of cosmetics, evaluation of eye irritancy caused by ingredients or metabolites is indispensable. Although the Draize rabbit eye irritation test (Draize test) proposed by Draize, Woodard, and Galvery (1944) has been conventionally used for this purpose, its use has been restricted recently due to mounting concerns about cruelty to animals. In the recent years, many methods that do not use animals have been developed (see, e.g., OECD 1996). One such method is a chorioallantoic membrane assay by trypan blue staining test (CAM-TB).

Yasutaka Chiba is a Doctorate Student, Department of Biostatistics, Kyoto University School of Public Health, Yoshida Konoe-cho, Sakyo-ku, Kyoto 606-8501, Japan (E-mail: chibay@pbh.med.kyoto-u.ac.jp). Yutaka Matsuyama is Associate Professor, Department of Biostatistics, Kyoto University School of Public Health. Tosiya Sato is Professor, Department of Biostatistics, Kyoto University School of Public Health. Isao Yoshimura is Professor, Department of Management Science, Tokyo University of Science, Kagurazaka 1-3, Shinjuku-ku, Tokyo 162-8601, Japan.

Journal of Agricultural, Biological, and Environmental Statistics, Volume 10, Number 1, Pages 118–130
DOI: 10.1198/108571105X28679

An interlaboratory study to validate alternative methods including CAM-TB to the Draize test was conducted. In the course of establishing a formula to predict Draize test scores based on an alternative method such as the CAM-TB, a statistical issue arose. One approach to establishing a prediction formula is to demonstrate a functional relationship between the scores in the Draize test and the measurements in an alternative method. Each of the paired measurements (the CAM-TB and the Draize test) was measured with error. These measurements were governed by heterogeneous errors. This kind of problem often arises in regression analysis with measurement error models (e.g., Fuller 1987) when error variances are heterogeneous between measurements. This article proposes a new method for analyzing these types of data.

In Section 2, the data obtained from the Draize test and the CAM-TB is described. In Section 3, we formulate the problem for analyzing these data and propose a statistical method to solve this problem. In Section 4, we conduct a simulation study to examine the performance of the proposed method in comparison with other methods, which do not assume heterogeneous error variances. The simulation study was conducted in a framework that was adaptable to the data obtained in our application. In Section 5, we analyze the data described in Section 2. Finally, we discuss the implications of the proposed method.

2. FORMULATION OF PROBLEM

The Draize test is a test designed to predict the eye irritation potential of chemical substances (Draize et al. 1944). The focus of this test is an assessment of observable mucosal and epithelial effects. Test scores represent damage caused to a rabbit's eyes on a scale from 0 to 110. See Wilhelmus (2001) for a more detailed explanation.

Several parties, including scientists and the Society for the Prevention of Cruelty to Animals, actively criticize the use of animal experiments based on the belief that experiments such as the Draize test are inhumane. Accordingly, there is growing demanding for alternative methods to animal experiments.

The CAM-TB is a test devised as an alternative method to the Draize test (Hagino et al. 1991, 1993, 1999). A chorioallantoic membrane (CAM) assay evaluates blood vessel reaction and damage to the CAM of a fertilized hen's egg. A hen's egg is recognized as experiment material that lies between in vivo and in vitro. Consequently, the CAM assay is considered to be less cruel than the Draize test.

In order to identify effective alternative methods, the Japan Cosmetic Industry Association (1994) conducted a validation study. In this study, the Draize test and the CAM-TB were carried out using 36 chemical substances. For each chemical substance, three rabbits and five eggs were examined. The results provided three Draize test scores and five CAM-TB measurements for each chemical substance.

Figure 1 shows a scatterplot of the scores in the Draize test by the measurements in the CAM-TB, where dots denote means and whiskers denote standard deviations. Figure 2 shows the mean SD plots, which show the association between means and standard deviations for the Draize test and the CAM-TB. In the Draize test, error variances of scores

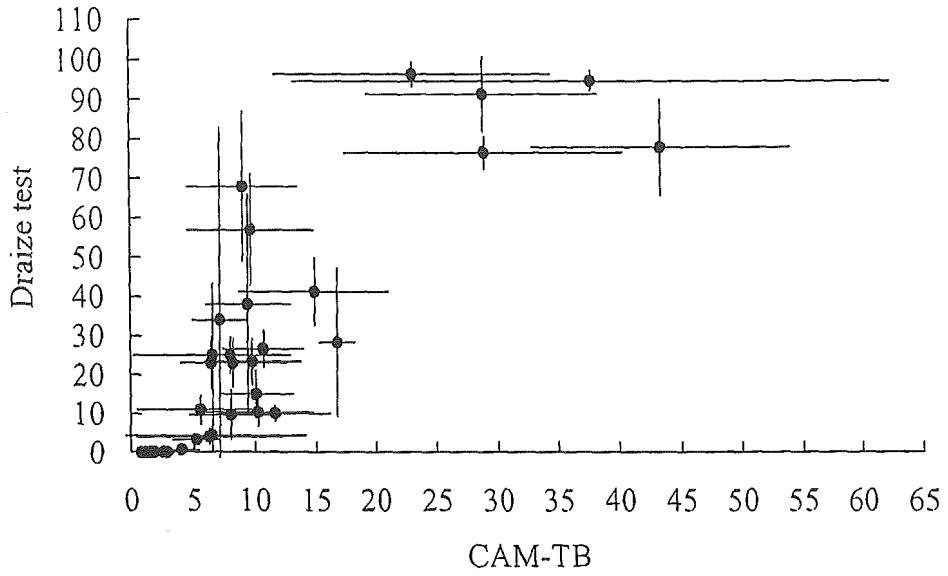


Figure 1. Scatterplot of the Draize Test and the CAM-TB: The vertical axis is the mean score of the Draize test and the horizontal axis is the mean of the CAM-TB, and whiskers express standard deviations.

for nonirritants or severe irritants were generally smaller than those for moderate irritants (top of Figure 2). In the CAM-TB, error variances of measurements increase in relation to the level of severity of irritants (bottom of Figure 2). Thus, we assume that error variance is dependent on the mean with a certain relationship. Because scores in the Draize test have an upper limit (110) and a lower limit (0), it is reasonable to fit a curve with sigmoid shape. Of all the curves with sigmoid shape, we choose the logistic curve because it is flexible and relatively easy to work with. We assume that the relationship between true scores in the Draize test, η, and true measurements in the CAM-TB, ξ, is well modeled by the logistic curve, $\eta = 110\exp(\beta_0 + \beta_1\xi) / \{1 + \exp(\beta_0 + \beta_1\xi)\}$. These actual data are measured with error. The principal problem of the validation study then lies in the evaluation of predictability of the CAM-TB to the Draize test. A suitable statistical method is required to estimate the regression parameters on a measurement error model with heterogeneous error variances. We employ the logit transformation, $\eta^* = \log\{\eta/(110 - \eta)\}$, which yields a linear relationship, to derive a suitable statistical method. This situation is formulated as follows.

Let $(x_i, z_i), i = 1, 2, \ldots, n$, denote a pair of measurements, and ξ_i, which is treated as a nuisance parameter, denote an unknown true value of x_i. A functional relationship exists between x_i's and z_i's such that

$$\begin{aligned} x_i &= \xi_i + \delta_i \\ z_i &= \beta_0 + \beta_1\xi_i + \varepsilon_i, \end{aligned}$$

where β_0 and β_1 are regression parameters, δ_i is the measurement error, and ε_i is the disturbance term in a linear regression model. It is assumed that δ_i and ε_i are distributed

according to

$$\begin{pmatrix} \delta_i \\ \varepsilon_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{xi}^2 & 0 \\ 0 & \sigma_{zi}^2 \end{pmatrix} \right).$$

It is necessary to estimate these variances.

3. PARAMETER ESTIMATION METHOD

In order to find a less-biased parameter estimation method for the model derived in Section 2, we considered expanding Amari and Kawanabe's proposal (1997). They

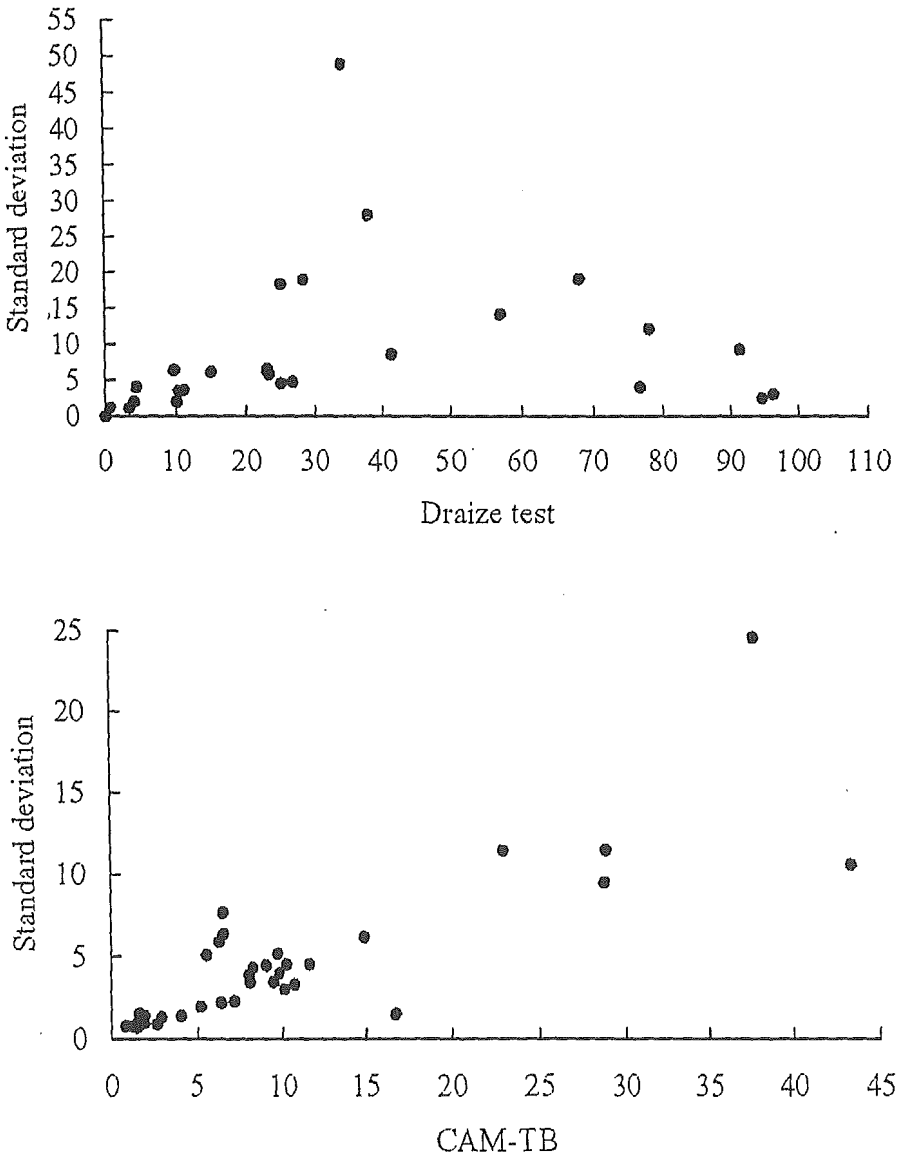


Figure 2. Top: Mean SD plot of the Draize test (vertical axis is standard deviation and horizontal axis is mean), Bottom: Mean SD plot of the CAM-TB (vertical axis is standard deviation and horizontal axis is mean).

discussed the problem that nuisance parameters increase with an increase in sample size. In some cases in which such problems arise, the maximum likelihood method does not derive consistent estimators (e.g., see Neyman and Scott 1948). The statistical analysis of linear measurement error models poses the same problem, that is, nuisance parameters increase with the increase of sample size. For this problem, Amari and Kawanabe (1997) defined a family of estimating functions under the assumption of homogeneous error variances. We extend their method to the case with heterogeneous error variances, and propose a parameter estimation method that makes asymptotic variances of parameters smallest in an extended family of estimating functions.

Under the common known variances, $\sigma_{xi}^2 = \sigma_{zi}^2 = \sigma^2$, Amari and Kawanabe (1997) showed that the estimating functions for β_0 and β_1, $g_0(x, z, \beta)$ and $g_1(x, z, \beta)$, become

$$g_0(x, z, \beta) = \frac{z - \beta_0 - \beta_1 x}{(1 + \beta_1^2)\sigma^2} \tag{3.1}$$

and

$$g_1(x, z, \beta) = h(s)\frac{z - \beta_0 - \beta_1 x}{(1 + \beta_1^2)\sigma^2}, \tag{3.2}$$

where $s = \{x + \beta_1(z - \beta_0)\}/(1 + \beta_1^2)$ is the sufficient statistic for ξ, and $h(s)$ is an arbitrary function of s. The estimating functions (3.1) and (3.2) can produce consistent estimators without depending on nuisance parameters $\xi_1, \ldots, \xi_n$. The asymptotic variance of $\beta_j, j = 0, 1$, a.v.$[\sqrt{n}(\hat{\beta}_j - \beta_j)]$, is given by

$$\text{a.v.}\left[\sqrt{n}\left(\hat{\beta}_j - \beta_j\right)\right] = \frac{\lim_{n\to\infty}(1/n)\sum_{i=1}^n E_{\beta,\xi_i}\left[g_j(x_i, z_i, \beta)^2\right]}{\left\{\lim_{n\to\infty}(1/n)\sum_{i=1}^n E_{\beta,\xi_i}\left[\partial g_j(x_i, z_i, \beta)/\partial\beta_j\right]\right\}^2}, \tag{3.3}$$

where $E_{\beta,\xi_i}[\cdot]$ denotes the expectation with respect to the distributions specified by β and ξ_i, and $\partial g_j(\cdot)/\partial\beta_j$ is the partial derivative with respect to β_j. The derivation of the asymptotic variance (3.3) is given in Appendix A.1.

We extend estimating functions (3.1) and (3.2) to the case with heterogeneous error variances. In order to simplify the estimating equation, we constrained $h(s)$ to a linear function, $h(s) = as + b$. The resulting estimating equations become

$$\sum_{i=1}^n \frac{z_i - \hat{\beta}_0 - \hat{\beta}_1 x_i}{\sigma_{zi}^2 + \hat{\beta}_1^2\sigma_{xi}^2} = 0 \Leftrightarrow \hat{\beta}_0 = \sum_{i=1}^n \frac{z_i - \hat{\beta}_1 x_i}{\sigma_{zi}^2 + \hat{\beta}_1^2\sigma_{xi}^2} \Big/ \sum_{i=1}^n \frac{1}{\sigma_{zi}^2 + \hat{\beta}_1^2\sigma_{xi}^2} \tag{3.4}$$

for β_0 and

$$\sum_{i=1}^n \left\{ a_i \frac{\sigma_{zi}^2 x_i + \hat{\beta}_1\sigma_{xi}^2(z_i - \hat{\beta}_0)}{\sigma_{zi}^2 + \hat{\beta}_1^2\sigma_{xi}^2} + b_i \right\} \frac{z_i - \hat{\beta}_0 - \hat{\beta}_1 x_i}{\sigma_{zi}^2 + \hat{\beta}_1^2\sigma_{xi}^2} = 0 \tag{3.5}$$

for β_1. It is necessary to decide on a_i and b_i to estimate $\hat{\beta}_1$, while $\hat{\beta}_0$ is uniquely determined. If we choose a_i and b_i to minimize the asymptotic variance (3.3), we get $a_i = 0$ and $b_i = \xi_i$. Equation (3.5) then becomes

$$\sum_{i=1}^n \xi_i \frac{z_i - \hat{\beta}_0 - \hat{\beta}_1 x_i}{\sigma_{zi}^2 + \hat{\beta}_1^2\sigma_{xi}^2} = 0. \tag{3.6}$$

$\hat{\beta}_1$ cannot be estimated by Equation (3.6) because it includes the unknown parameter ξ_i. Accordingly, we estimate $\hat{\beta}_1$ by substituting the maximum likelihood estimator $\hat{\xi}_i = \left\{\sigma_{zi}^2 x_i + \hat{\beta}_1 \sigma_{xi}^2 (z_i - \hat{\beta}_0)\right\} / (\sigma_{zi}^2 + \hat{\beta}_1^2 \sigma_{xi}^2)$ of ξ_i into Equation (3.6). This $\hat{\xi}_i$ is consistent with the sufficient statistic s_i for ξ_i. As the result, $a_i = 0$ and $b_i = \xi_i$ derive $h(s_i) = s_i$. The estimating equation for β_1 becomes

$$\sum_{i=1}^{n} \frac{\left\{\sigma_{zi}^2 x_i + \hat{\beta}_1 \sigma_{xi}^2 (z_i - \hat{\beta}_0)\right\} (z_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)}{(\sigma_{zi}^2 + \hat{\beta}_1^2 \sigma_{xi}^2)^2}. \tag{3.7}$$

Incidentally, $\hat{\beta}_0$ and $\hat{\beta}_1$ in Equations (3.4) and (3.7) are the same as the maximum likelihood estimators (Walter 1997). The derivations of Equations (3.4) and (3.7) through the maximum likelihood method are given in Appendix A.2. The variances of $\hat{\beta}_0$ and $\hat{\beta}_1$ are given by

$$\begin{aligned} \mathrm{var}(\hat{\beta}_0) &= \frac{\sum f_i \xi_i^2}{\sum f_i \sum f_i \xi_i^2 - \left(\sum f_i \xi_i\right)^2} \\ \mathrm{var}(\hat{\beta}_1) &= \frac{\sum f_i}{\sum f_i \sum f_i \xi_i^2 - \left(\sum f_i \xi_i\right)^2}, \end{aligned}$$

where $f_i = 1/(\sigma_{zi}^2 + \beta_1^2 \sigma_{xi}^2)$ and $\sum = \sum_{i=1}^{n}$.

4. SIMULATION STUDY

We conducted a simulation study to evaluate the performance of the parameter estimation method proposed in Section 3. The simulation study was conducted in a framework that was adaptable to data obtained in a validation study described in Section 2. The following four parameter estimation methods were compared:

1. Ordinary least squares (we call this method OLS).
2. Estimating Equations (3.4) and (3.7) under the assumption of homogeneous error variances, and the error variance of response variables are equal to that of explanatory variables, that is, $\sigma_{xi}^2 = \sigma_{zi}^2 = \sigma^2$ (we call this method EV1).
3. Estimating Equations (3.4) and (3.7) under the assumption of homogeneous error variances, and the error variance of response variables are not equal to that of explanatory variables, that is, $\sigma_{xi}^2 = \sigma_x^2$ and $\sigma_{zi}^2 = \sigma_z^2$ (we call this method EV2).
4. Estimating Equations (3.4) and (3.7) under the assumption of heterogeneous error variances, and all error variances of response variables and explanatory variables vary between measurements (we call this method NEV).

We substituted the Equation (3.4) into Equation (3.7) to find $\hat{\beta}_1$ in EV1, EV2, and NEV. The Newton-Raphson algorithm was used to solve estimating equations. The initial values were set to be OLS estimates. The conditions and steps for the simulation study were as follows:

1. Generate a uniform random number with a range between 5 and 15. This value is set as ξ, which denotes a true measurement in the CAM-TB.

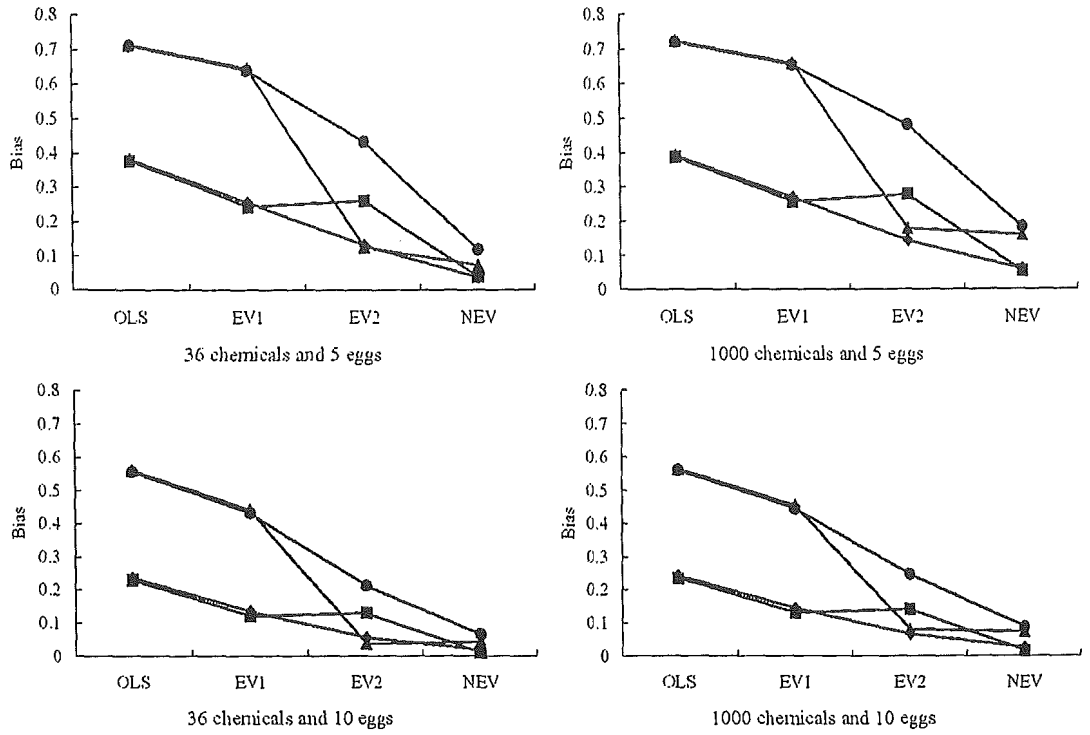


Figure 3. The Results of $\hat{\beta}_1$ for a Simulation Study: Biases (true values minus the means of estimates). Diamonds: $(SD_x, SD_y) = (.5\xi, .0025\,\eta(110-\eta))$, squares: $(SD_x, SD_y) = (.5\xi, .005\,\eta(110-\eta))$, triangles: $(SD_x, SD_y) = (\xi, .0025\,\eta(110-\eta))$, and circles: $(SD_x, SD_y) = (\xi, .005\,\eta(110-\eta))$.

2. Calculate $\eta = 110\exp(\xi - 10)/\{1 + \exp(\xi - 10)\}$ (true model). This η denotes a true score in the Draize test.
3. Generate normal random numbers with mean ξ and standard deviation $\mathrm{SD}_x = .5\,\xi$ or ξ. The number of generated random numbers is 5 or 10, which is assumed to be the number of eggs used in the CAM-TB. Set the mean and standard deviation calculated from these random numbers as a measurement and its according standard deviation in the CAM-TB. Similarly, generate normal random numbers with mean η and standard deviation $\mathrm{SD}_y = .0025\eta(110-\eta)$ or $.005\eta(110-\eta)$. If a generated value is smaller than 0 or larger than 110, the value is replaced with 0 or 110. The number of generated random numbers is 3, which is the same as the number of rabbits used in the Draize test. Set the mean and standard deviation calculated from these random numbers as the score and its according standard deviation in the Draize test.
4. Repeat Steps 1–3 36 times (in correlation to the number of chemical substances) or 1,000 times. This yields one dataset.
5. Estimate parameters (β_0 and β_1) by the four methods described (OLS, EV1, EV2, and NEV).
6. Repeat the above steps 1,000 times, and calculate means, biases (true values minus means), and mean squared errors (MSEs) of parameter estimates ($\hat{\beta}_0$ and $\hat{\beta}_1$).

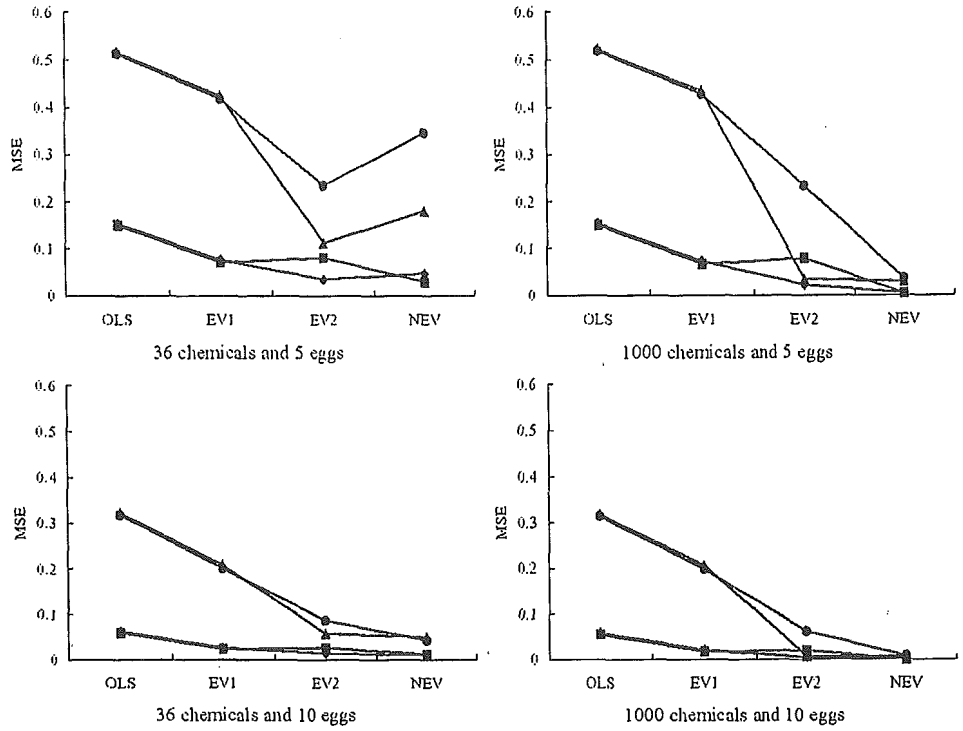


Figure 4. The Results of $\hat{\beta}_1$ for a Simulation Study: Mean squared errors. Diamonds: $(SD_x, SD_y) = (.5\xi, .0025\,\eta(110-\eta))$, squares: $(SD_x, SD_y) = (.5\xi, .005\,\eta(110-\eta))$, triangles: $(SD_x, SD_y) = (\xi, .0025\,\eta(110-\eta))$, and circles: $(SD_x, SD_y) = (\xi, .005\,\eta(110-\eta))$.

The results of the simulation study are summarized in Figures 3 and 4. We show only the results of $\hat{\beta}_1$ since $\hat{\beta}_0$ was uniquely determined as a result of $\hat{\beta}_1$. Figure 3 indicates biases on the four methods under $4 \times 2 \times 2 = 16$ conditions (pairs of standard deviations $(\mathrm{SD}_x, \mathrm{SD}_y) = (.5\xi, .0025\eta(110-\eta))$, $(.5\xi, .005\eta(110-\eta))$, $(\xi, .0025\eta(110-\eta))$ or $(\xi, .005\eta(110-\eta))$; the number of chemical substances = 36 or 1,000; and the number of eggs = 5 or 10). Figure 4 indicates mean squared errors.

The following results were obtained from Figure 3. First, magnitude of biases for the four methods was examined. All four results in Figure 3 show that NEV, EV2, EV1, and OLS generated smaller order biases. However, the biases of EV2 are a little larger than those of EV1 under $(\mathrm{SD}_x, \mathrm{SD}_y) = (.5\xi, .005\eta(110-\eta))$ (symbols in Figure 3 = squares). The biases of EV2 are also similar to those of NEV under $(\mathrm{SD}_x, \mathrm{SD}_y) = (\xi, .0025\eta(110-\eta))$ (symbols in Figure 3 = triangles).

Second, we examined how the difference of standard deviations (SD_x or SD_y) influenced biases. All four results in Figure 3 show that biases for $\mathrm{SD}_x = \xi$ (symbols in Figure 3 = triangles and circles) are larger than those for $\mathrm{SD}_x = .5\xi$ (symbols in Figure 3 = diamonds and squares). On the other hand, the difference of SD_y, that is, $\mathrm{SD}_y = .0025\eta(110-\eta)$ (symbols in Figure 3 = diamonds and triangles) or $\mathrm{SD}_y = .005\eta(110-\eta)$ (symbols in Figure 3 = squares and circles), does not generate much difference in biases except for EV2. These results imply that error variances in the CAM-TB have more influence on biases than those of the Draize test. This is especially true for OLS and EV1. It seems that