Figure 6.4. MEASURING: Three Examples from Physics

Much ingenuity and labour has been expended on measurement of the values of physical constants; whenever *independent* measurements have been made, a critical appraisal of the data is needed to pick a 'best' value for the constant and to set some limit for the uncertainty (or 'error') of this value. The major theme of this Figure is that much *more* can be done by experimenters than has been done in the past to assist both the experimenter and the critical appraiser in their tasks, particularly with regard to the *proper* evaluation of inaccuracy (*i.e.*, systematic error).

Measurement error is usually considered to be the net result of several components:

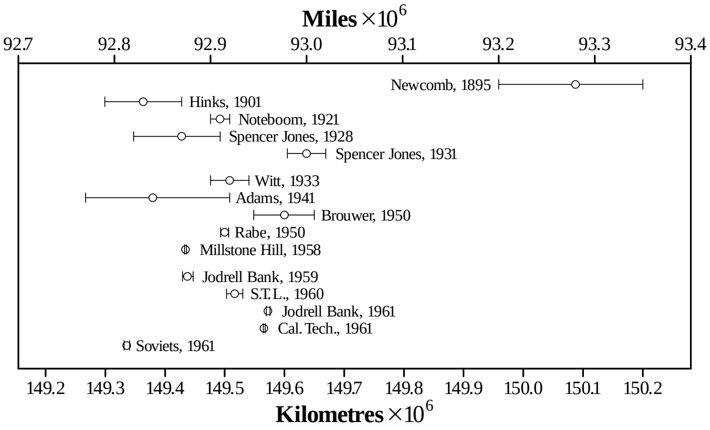
- *imprecision*: variation among repeat results with no deliberate change in the apparatus or procedure;
- *inaccuracy* (or *systematic error*)
  - associated with the theory of the method of measurement;
  - resulting from the apparatus or observer.

Imprecision can be reduced and quantified by repetition and averaging and is usually minor; the role of inaccuracy is apparently underestimated, as illustrated in the three diagrams at the right.

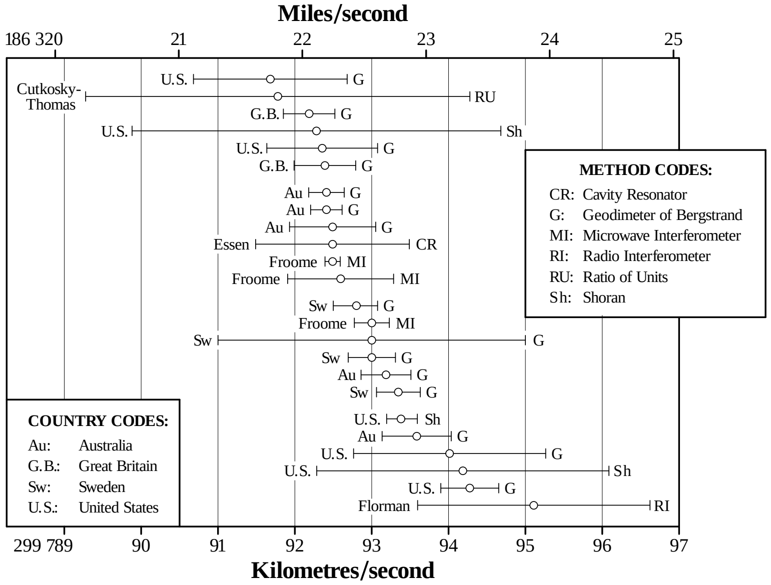
- Diagram (a) shows the results from 15 determinations of the astronomical unit – the average distance between the earth and the sun; the centre of each circle (○) gives the 'best' (point) estimate from the relevant investigation, and the bars (⊢—⊣) indicate the investigators' assessment of the limits of error in their values. Each bar is labelled with the name(s) of the investigator(s) and the year of their report of the work.
- Diagram (b) shows the results from 24 determinations of the speed of light, using the same conventions as in diagram (a), except that for non-laboratory measurements, the *country* rather than the investigator(s) is given; in addition, a code for the *experimental method* for the determination appears at the right-hand end of each error bar. The studies cover a period of about 30 years and the results are ordered according to the value obtained.
- Diagram (c) shows the results from a Canadian determination of the acceleration due to gravity at Ottawa. The method involved timing the drop of a rule in an evacuated tube; results were reported for 32 drops of each of two equally carefully calibrated rules. Each set of results is shown as a histogram, the numbers within the bars being the last two digits of the measurements (*e.g.*, '06' represents 980.6106 cm/sec^2). The circle and bar under each histogram represent the average ± one standard deviation of the 32 values.

REFERENCE: Youden W. J.: Enduring Values, *Technometrics* **14**(18): 1-11 (1972).

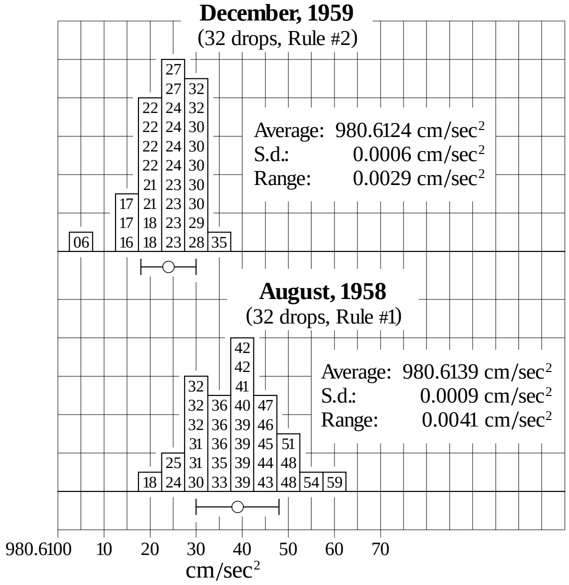
(a) The Astronomical Unit



(b) The Speed of Light



(c) The Acceleration due to Gravity at Ottawa



(*continued overleaf*)

A noteworthy feature of the results in diagram (a) is that each investigator's 'best' value lies *outside* the limits reported by his immediate predecessor, and the situation is not greatly improved even if the *ranges* of the error bars are considered instead of the point estimates; this suggests these investigators do *not* have reliable information about their sources of inaccuracy. In diagram (b), the values are spread over a range of 3.5 km/second yet half the claimed errors are well under 0.5 km/second, suggesting that individual investigators are unable to set realistic limits of error for their reported values. In diagram (c), despite the equally careful calibration of the two rules, they clearly lead to different results, providing us with a glimpse of how inaccuracy is connected with some part of the apparatus. The lesson is that an *essential* responsibility of all experimenters is to try to *quantify* this and other types of inaccuracy. As discussed below, by carrying out an experiment in a properly designed manner, inaccuarcy from different sources can be estimated *efficiently* (*i.e.*, without having to make an excessive number of measurements) and *experimentally* (*i.e.*, without relying merely on 'paper' estimates based on expert judgement).

Suppose we are concerned with quantifying *three* sources of measuring inaccuracy in the determination of the acceleration due to gravity: the calibration of the rule [which the data in diagram (c) overleaf on page 6.7 address], and the timing clock and the residual pressure in the tube [which these data do *not* address]. Denote the two rules by R_1 and R_2, two clocks by A and B and two pressures by P_1 and P_2; then the *same* number of measurements (64) as in diagram (c) should be made comprising 16 with each of the following four combinations of the three sources of inaccuracy (rule, clock, pressure): (1) R_1AP_1 (2) R_1BP_2 (3) R_2BP_1 (4) R_2AP_2. The comparison of rules is then made by comparing (1) + (2) with (3) + (4); the comparison is *fair* in that it is not contaminated with clock or pressure inaccuracy and, being based on all 64 measurements, is of the *same* precision as for the data in diagram (c). Further, the equivalent comparison of *clocks* now comes from (1) + (4) *vs* (2) + (3) and that of *pressures* from (1) + (3) *vs* (2) + (4).

The basis of the experimental design described in the preceding paragraph, in which the *same* amount of data yields essentially *three times* as much information as diagram (c), is the *systematic changing of more than one factor at a time* in different experimental 'runs'. This is contrary to most people's intuition, which instead generally leads to the inefficient practice of changing *only one factor at a time*. For example, if the four runs are based on (1) R_1AP_1 (2) R_2AP_1 (3) R_1BP_1 (4) R_1AP_2, only *half* the data are used in making each comparison, resulting in an appreciable loss of sensitivity. Another disadvantage of the 'change one variable at a time' approach is that a simple average of these four combinations does *not* give equal weight to the alternatives. These matters illustrate important statistical principles: *data generated according to a proper experimental design are usually more informative* and *easier to analyse*.

In his concluding paragraphs, Youden says: What seems most interesting to me is that when measuring a physical constant usually no changes are made in the variables, or at most in one or two, as in the Ottawa study. Equally clear is the fact that when another investigator makes his effort, *every* component in the measuring system is changed. Reliance is again placed upon calibration and various corrections. Nothing is done about the fact that the results obtained by different investigators disagree a great deal more than would be expected in view of the 'paper' estimates of systematic error used by the investigators. It appears to be an all or none situation. Everything gets changed in another laboratory. Almost nothing gets changed within a laboratory. This makes it impossible to single out and measure the sources of measuring inaccuracy. The expert faced with finding a best value is equally in the dark. His role is more like the man judging animals at a county fair than that of a scientist. And it would be so easy to improve on this state of affairs.

$\boxed{1}$ In the discussion on page 6.7 overleaf taken from Youden, three *sources* of measuring inaccuracy are mentioned. What are they?

- For what sources, *other than* measuring, do data become inaccurate?

$\boxed{2}$ Diagram (b) presents the results of a sizeable collection of serious attempts to measure the speed of light. From the information in the diagram, explain briefly what do you conclude about:

- the 'best' value (or point estimate);
- the limits of error of this estimate.

$\boxed{3}$ In diagram (b) overleaf on page 6.7, the 15 values based on the geodimeter method (code 'G') are spread approximately uniformly over the whole range of the 24 values, and the three Shoran ('Sh') results are *4*th, *19*th and *22*nd in order of magnitude. What conclusion do you draw about inaccuracy associated with the *theory* of these two methods of measuring the speed of light?

$\boxed{4}$ In diagram (c) overleaf on page 6.7, explain briefly why the difference in the averages of the two sets of 32 measurements is more likely due to *inaccuracy* than to *imprecision*.

$\boxed{5}$ In light of the diagrams overleaf on page 6.7 and the discussion above, what *reason(s)* do you infer for the fact that different investigators disagree by more, sometimes *far* more, than their claimed errors?

- What do you conclude about 'paper' estimates of inaccuracy based on expert judgement? Explain briefly.

$\boxed{6}$ Explain briefly the meaning of the following statements, which occur in the second and third paragraphs above:

- *the comparison is* fair *in that it is not contaminated with clock or pressure inaccuracy*;
- *a simple average of these four combinations does* not *give equal weight to the alternatives.*

$\boxed{7}$ Explain what Youden has in mind in his concluding statement: *And it would be so easy to improve on this state of affairs.*