Archives of Disease in Childhood 1990; 65: 1340–1344

The reliability of height measurement (The Wessex Growth Study)

L D Voss, B J R Bailey, K Cumming, T J Wilkin, P R Betts

Abstract

The two major components of reliability are accuracy and reproducibility. Three studies of the reliability of height measurement in children are reported. In the first, a standard metre rod was used to spot check the accuracy of installation of 230 measuring instruments in one health district in Wessex, UK. The readings obtained ranged from 90·0 to 108·5 cm and showed the urgent need for the positioning of instruments to be regularly checked.

In a second study, to examine the reproducibility of height measurement, two experienced observers measured 10 young children (106·0 to 152·0 cm), three times on five instruments of different design. The observations were blind and in random order. The estimated standard deviation for a single height measurement was generally in the range 0·2–0·3 cm. Over 95% of the variance was attributable to the child, very little to the instrument or observer.

Finally, the conditions of the second study were modified to examine the effect on reproducibility of non-blind and non-randomised measurements, as usually occurs in the clinic. A lower but inevitably false estimate of the error was obtained. It is recommended that the error of height measurement, appropriately established and expressed in simple terms, be stated in every published growth study.

Quality control data, although routinely reported in the laboratory, is conspicuously absent from studies relating to growth. It is presumably felt that the measurement of height is a simple task and, provided sufficient care is taken, any error will be so small that it can be safely ignored. Hindmarsh and Brook maintain that the measurement of height can be 'extremely accurate',[1] but Tanner warns that the 'heights measured in the averagely casual clinic are useless even for accurate clinical purposes, let alone research'.[2]

Earlier this century, height measuring apparatus was necessarily crude by today's standards—the instrument recommended for use in schools was Baldwin's paper measuring scale,[3] and researchers were well aware of the variability of anthropometric data. Krogman has reviewed many studies which attempted to quantify the error.[4] Instruments today are infinitely more sophisticated, which may be why many recent growth studies make no reference to any error of measurement at all.[5–8]

There is error, however, in all measurement and the validity of growth data, both in the clinic and in research, depends critically on the reliability of height measurements. Attempts of modern researchers to establish their error of measurement have often been unsatisfactory, for two reasons. First, the conditions under which the observations were made have been inadequately described and cannot therefore be replicated by others.[9 10] Second, as Cameron pointed out, there is no commonly agreed terminology to express the error of height measurement, making cross study comparisons unnecessarily difficult.[11]

This error in height measurement has been variously expressed as coefficient of variation (CV),[1] standard deviation (SD),[9] standard error of measurement (S_{meas}),[12] or standard error of the mean (SEM).[13] The coefficient of variation is defined as $SD/\bar{x} \times 100\%$, where $\bar{x}$ is the mean of the observations. The SD of a height measurement should not vary greatly with height itself, so that as a child grows, the coefficient of variation will become smaller. The error as expressed by the coefficient of variation therefore appears to diminish when measuring taller children. In practice, however, the error may be just as large and therefore just as critical for monitoring velocity, where height increment, not absolute height, is important. SD and standard error of measurement are synonymous and can be calculated from the differences observed between two or more measurements. The SD, once established, can be applied to a *single* future measurement made under similar conditions. Where the mean of *several* measurements is used in order to reduce variability, this SD can be divided by $\sqrt{n}$ to give the standard error of the mean. (Reference to an SEM would be quite inappropriate where only *single* observations are to be made, as is commonly the case, and would give a misleadingly low estimate of the error.) This study uses the SD for a single height measurement to express the error.

The aims of the present study were to establish the reliability of height measurement by examining: (i) The accuracy of installation of height measuring instruments. (ii) The reproducibility of height measuring instruments in a rigourously controlled trial, quantifying the error in clear and simple terms and analysing the contribution made to the total variance by instrument, observer and subject. (iii) The effect on the error of varying measurement conditions.

The definitions used are:

Reliability—The reliability of growth data depends on both accuracy and reproducibility

Southampton General Hospital, Endocrine Section, Medicine II
L D Voss
T J Wilkin

Paediatric Medical Unit
K Cumming
P R Betts

Faculty of Mathematical Studies, University of Southampton
B J R Bailey

Correspondence to:
Ms L D Voss,
(Endocrine Section)
Medicine II,
Level D, South Block,
Southampton General Hospital,
Southampton SO9 4XY.

Accepted 7 June 1990

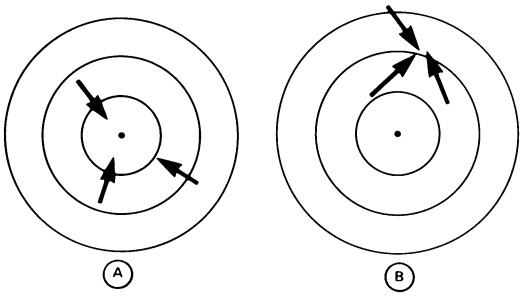

Figure 1 Dartboard analogy to illustrate the components of reliability: accuracy and reproducibility. (A) Observations are accurate but reproducibility is poor. (B) Reproducibility is good but observations are inaccurate.

(fig 1). These are quite different, yet the terms are sometimes confused.

Accuracy—This is a measure of the closeness of observations to the target, in this case the 'true' height. With an accurate instrument, the mean of a large number of observations would hit the target, irrespective of the size of their spread, provided the error is random (fig 1A). Badly installed equipment will introduce a systematic error, leading to inaccurate measurements.

Reproducibility (precision)—Repeat measurements on the same child often differ. Observations will be distributed around their mean, with some spread that could be expressed as, for example, the variance. The smaller the variance the greater the reproducibility. Measurements would be reproducible without being accurate, if the spread about their mean were small, but that mean did not coincide with the target value (fig 1B).

Sources of variance—The instrument, observer, and subject are all sources of error, and the variance of each contributes to the total variance of measurement.

Methods

(1) INSTALLATION OF HEIGHT MEASURING INSTRUMENTS

An aluminium metre rod was used by a single observer (KC) to spot check the accuracy of installation of 17 stadiometers, 55 Microtoises, 133 rulers, and 25 wall charts in everyday use in health centres, hospitals, schools, and general practice surgeries in Wessex. No prior warning was given to personnel in charge of the equipment. The recorded length of the rod was based on the mean of two readings.

The instruments examined fell into one of four groups:

Stadiometers (Holtain Ltd, Crymmych, Wales)—The 'Harpenden' model is the standard instrument in paediatric outpatient departments in the UK. It consists of a vertical backboard with a weighted horizontal cursor fixed at 90° to it. A mechanical counter running on a track records the height.

Microtoises (CMS Weighing Equipment Ltd)—A metal tape, hung from a permanent hook on the wall, is pulled down onto the child's head. It is used predominantly by peripatetic school nurses.

Ruler—Any simple vertical scale with a moveable horizontal bar was described as a ruler. These are frequently found in general practice and in schools.

Wall charts—These included any charts attached to the wall against which the child stands. A right angled block is placed on the child's head and the height read directly off the chart. They were found mainly in general practice and in health centres.

(2) REPRODUCIBILITY OF HEIGHT MEASUREMENTS—STANDARD EXPERIMENTAL CONDITIONS

Instrument and observer alone

Two blocks of machine cut wood, of different but unknown length, were measured 10 times each by a single observer on five different instruments: (a) *Holtain electronic stadiometer*—the headpiece, when horizontal, activates a light beam which then reads the height off a binary scale giving a digital readout; (b) *Harpenden stadiometer*—as described earlier; (c) *Raven Magnimetre*—the fixed backboard has a removable magnetic measuring arm; (d) *Raven Minimetre*—similar to the Microtoise, but more easily repositioned as it can be stuck to the wall with plastic adhesive; and (e) *Harpenden pocket stadiometer*—the head bar with retractable metal tape is hooked to a freestanding base plate. It is difficult to ensure the tape is taut while keeping the head bar horizontal.

The order of measurements was randomised and they were also 'blind', so that although the observer positioned the blocks and lowered the headpiece, a third party noted and recorded the readings to the nearest 1 mm. The instruments were checked for accuracy using a metre rod both at the start and at the end of the trial and were found to be unchanged.

Instrument, observer, and child

Ten children (aged 4 to 11 years, heights 106·0 to 152·0 cm) were each measured three times by two experienced observers on the five accurately installed instruments described above. These measurements were also made under standard experimental conditions—that is, blind and in random order. Conventional anthropometric methods were used throughout.[11]

(3) REPRODUCIBILITY OF HEIGHT MEASUREMENTS—VARYING CONDITIONS

In a further trial, 10 children (eight from the previous trial) were measured by two experienced observers (one from the previous trial), seven times each using just two instruments: the Harpenden stadiometer and the Raven Magnimetre. The trial was designed in such a way as to allow a comparison of the reproducibility of height measurements under three different conditions: (i) Blind, random order: the standard conditions described above were repeated. (ii) Blind, successive: the observers again measured each child 'blind', but three times in quick succession. The children were removed from the instruments between measurements but immediately replaced so that no other children were measured between. (iii) Non-blind, suc-

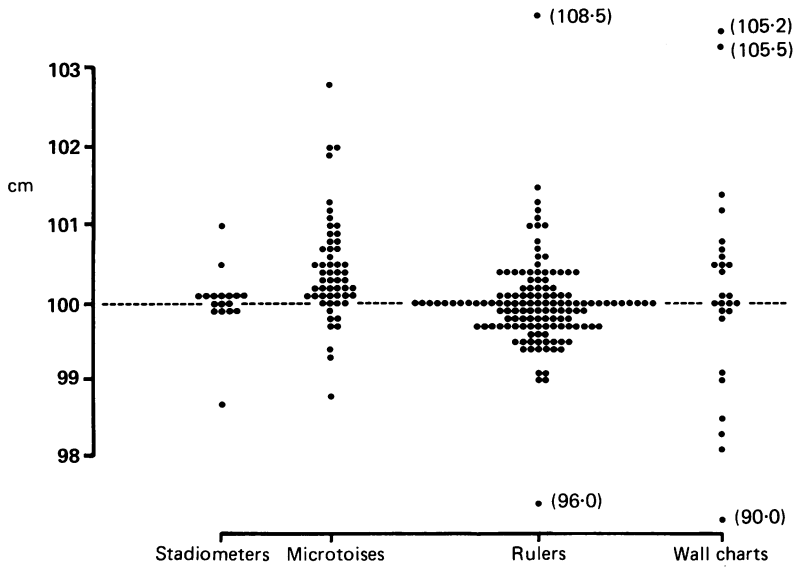

Figure 2 The length of a 100 cm metal rod as recorded by 230 measuring instruments installed in clinics throughout Wessex.

cessive: the children were measured three times in quick succession, and the measurements were no longer blind. Each observer was asked to read and record his own data.

Results

(1) INSTALLATION OF HEIGHT MEASURING INSTRUMENTS

Figure 2 shows the individual results of measuring the same metre rod, grouped according to instrument type. There were many cases of badly installed equipment. In each group there was an error of at least 1 cm above or below 100 cm; for some types of instrument, it was considerably greater.

The range of error was least for stadiometers (98·7 to 101·1 cm) and greatest for wall charts (90·0 to 105·2 cm). In the case of Microtoises, rulers, and wall charts, inaccuracies arose from positioning the instrument at the wrong height. The mean reading for the Microtoises, significantly greater than 100·0 cm ($p<0·001$), suggested a systematic tendency for these instruments to be placed too low. Incomplete extension of the tape during installation is the most likely cause. Error in the stadiometers was mainly due to incorrect setting of the digital counter on its track as a result of rough handling. Personnel using these instruments were almost without exception unaware of the inaccuracies.

(2) REPRODUCIBILITY OF HEIGHT MEASUREMENTS —STANDARD EXPERIMENTAL CONDITIONS

Instrument and observer alone

Table 1(a) shows the reproducibility of each instrument when measuring a rigid block of wood. The SD of a single height measurement was calculated. This ranged from 0·0 cm (Holtain electronic stadiometer, whose digital readout measures only to the nearest mm) to 0·1 cm (Raven Minimetre and Harpenden pocket stadiometer, where the variability, though greater, was still very small). The accuracy of each instrument, as measured by a metre rod, did not vary throughout the trial.

Table 1 Reproducibility of height measurements (standard experimental conditions)

Instrument	*(a) Instrument and observer alone SD (cm)*	*(b) Instrument, observer, and child SD (cm)*
Holtain electronic stadiometer	0·00	0·29
Harpenden stadiometer	0·07	0·26
Raven Magnimetre	0·03	0·17
Raven Minimetre	0·10	0·23
Harpenden pocket stadiometer	0·10	0·58

SDs are based on measurement of (a) two wooden blocks of unknown length and (b) 10 children, pooling data from two observers measuring each child three times on each instrument.

Instrument, observer, and child

Table 1(b) shows the reproducibility of each instrument when measuring children as opposed to wooden blocks. The SD is much greater, ranging from 0·17 cm (Magnimetre) to 0·58 cm (pocket stadiometer) when averaged over children and observers. This last instrument was significantly less reproducible when measuring children than the other four, among which there was no statistically significant difference in reproducibility ($p=0·14$). Neither was there any significant difference between the reproducibility of measurements made by the two observers, on any of the five instruments (p values ranged from 0·075 to 0·90), allowing their data to be pooled in table 1(b). The observers did, however, differ in technique and it is clear from table 2 that there was a significant difference between the two observers in mean height when using the Holtain electronic stadiometer, the Harpenden stadiometer, and the Raven Magnimetre. Where both hands were free to stretch a child, observer Y measured children significantly taller than X for all three instruments. It also follows that the height of a child may differ according to the technique involved in using different instruments. Pooling the data for the two observers, mean heights of the children were greater using the Holtain electronic stadiometer and Raven Magnimetre than when using the Raven Minimetre and Harpenden pocket stadiometer where no stretching was possible.

There was a pronounced difference in the reproducibility of height measurement between individual subjects. Looking only at the data for the Holtain electronic stadiometer, Harpenden stadiometer, and the Raven Magnimetre, where techniques of measurement were similar, a wide range of observed SDs for the 10 children was nevertheless evident, ranging from approximately 0·1 to 0·4 cm. Analysis of the components of variance showed that the subjects' contribution to the total variance ranged from 100% (electronic stadiometer) to 88% (Minimetre), observers and instruments accounting for the remainder.

(3) REPRODUCIBILITY OF HEIGHT MEASUREMENTS —VARYING CONDITIONS

Table 3 shows the SDs obtained in a further

Table 2 Mean heights (cm) recorded by different observers

	Holtain electronic stadiometer	*Harpenden stadiometer*	*Raven Magnimetre*	*Raven Minimetre*	*Harpenden pocket stadiometer*
Observer X	130·10	129·76	130·06	129·99	129·94
Observer Y	130·28	130·09	130·27	129·91	129·99
Difference X-Y	−0·18	−0·33	−0·21	0·08	−0·05
p Value	0·027	0·00014	0·00012	0·19	0·76

Mean heights shown are for 10 children measured by two observers three times each on the five instruments.

Table 3 Reproducibility of height measurements (varying conditions)

	(i) SD (cm)	*(ii) SD (cm)*	*(iii) SD (cm)*
Harpenden stadiometer	0·36	0·22	0·11
Raven Magnimetre	0·21	0·19	0·16

SDs are based on measurements made under the following conditions (i) standard experimental—that is, blind and in random order, (ii) blind, successive, and (iii) non-blind, successive.

trial under the following three conditions: (i) standard experimental—that is, blind, randomised, (ii) blind, successive, and (iii) non-blind, successive. The SDs under standard conditions (i) were comparable with those obtained in the previous trial. The SDs were smaller however under conditions (ii) than (i), and further reduced under (iii), very clearly so as in the case of the Harpenden stadiometer.

Discussion

Our data should give rise to concern, whether in the community, hospital clinic, or research department. Correctly installed instruments have been shown to measure a metre rod or wooden block with good reproducibility. Screening for abnormal stature in the community, however, may be liable to serious inaccuracy through malpositioning of the measuring instruments. As the school entry medical is often the first and only time the height of a child is formally assessed, abnormally small children missed then might not be referred for a specialist opinion until a much later age, by which time the potential for modifying the final adult height may be greatly reduced. The installation of height measuring instruments can be easily and quickly checked with a metre rod at the beginning and end of each session. We have shown that, accurately installed and correctly used, an inexpensive Microtoise or Minimetre may be no less reliable than a more expensive instrument.

Where children are measured more than once in order to monitor their growth, reproducibility of height measurement is critical. Even under ideal conditions—that is, blind and randomised—the SD of a single height measurement was fairly constant, generally between 0·2 and 0·3 cm, the range previously observed by Tanner[13] and by the present authors in routine quality control checks.

Instruments themselves appear to contribute very little to the total variability and attempts to design still more sophisticated models may not be worth the effort. There may, of course, be an interaction between instrument and subject that is not apparent when only a wooden block is measured. For example, while the pocket stadiometer was reproducible when measuring the wooden block, it performed less well measuring a child, probably because the design incorporates no fixed horizontal or vertical parts.

Trained and experienced observers differ little among themselves in reproducibility of their measurements. We have shown, however, that differences in technique call for repeat measurements on individual children to be made not only with the same instrument, but also by the same observer. The difference in heights recorded by different observers also implies that either one or both of the observers is not measuring the 'true' height of the child, that is, the measurements are biased *or* there is no such thing as the true height of a child.

The greatest source of variablity was clearly the child himself. The ideal child to measure would be rigid. A living subject, however, is of no fixed height. Independently of diurnal variation, posture can vary from one moment to the next, and the aim of measurement can only be to estimate a child's mean height by making several separate observations.

The reproducibility of *individual* subjects on *individual* instruments varied considerably. Accordingly, statements of the error of measurement based on selected cases could be misleadingly low. It is of practical use to know how small the error *can* be. What is required, if there is no time to calibrate each individual child, is the best estimate of the likely error, given a child presumed to be 'average'. Hence the value of an 'estimated standard deviation', established for that particular combination of observer and instrument, as a clear and simple expression of error.

We have shown that the SDs of successive and non-blind observations tend to be lower than those obtained under standard experimental conditions, giving an estimate of the error that is artificially small. Even where children are repositioned and observations are blind, successive measurements, where no other children are measured between, underestimate the variability. Under such conditions, the observer is perhaps better able to replicate his technique, or the child to retain the same posture. Posture might be expected to change throughout the day in much the same way as, for example, the pulse rate. The pulse rate sampled randomly over a long period would have an accurate mean but large variance. Sampled over a short period, however, successive observations would have a smaller variance, but a mean that may not necessarily represent the 'usual' rate for the subject.

Our data point to the need for measurements

to be not only random but blind as well. Where the observer is aware of the first measurement, he may consciously or otherwise try to make the next and subsequent readings as close as possible. Only the first reading is unbiased; the others have distributions conditional on the first, and the smaller SDs obtained relate only to the internal variability between the correlated measurements, not to their scatter around the target or 'true' height.

A similar problem to this arose many years ago with the need for reliable, uncorrelated measurements of blood pressure in drug trials.[14] It was resolved by developing the random zero sphygmomanometer, an instrument with which clinicians are able to perform repeat 'blind' measurements unaided, as the previous reading is disguised. A random zero stadiometer would be a useful tool in auxology.

Non-blind and/or successive observations could explain some of the high reproducibility of measurement claimed by other authors. Accuracy, or closeness to the target (fig 1A) should be the goal, however, and should not be sacrificed for high reproducibility (fig 1B). If confidence intervals are to be attached to a height measurement, an estimate of the SD of the larger variance about the mean in fig 1A is needed, not the smaller internal variance in fig 1B.

CONCLUSIONS

(1) The measurement of height is seriously hampered by the inaccurate installation of measuring instruments. (2) Some error is inevitable in measuring the height of children, the child himself being the major source of variance. Even where a trained observer uses an accurate and reproducible instrument, this error persists and the estimated SD is remarkably constant, generally in the region of 0·2 to 0·3 cm for a single height measurement. (3) Smaller reported errors could arise through failure to observe standard experimental conditions, or the reporting only of selected cases.

RECOMMENDATIONS

(1) Height measuring instruments should be regularly checked with a metre rod to ensure they are accurately set up. (2) All personnel concerned with the measurement of height should establish their own SD for a single height measurement, using their own instruments and a representative sample of subjects in a 'blind' and randomised calibration trial. (3) The monitoring of height should be carried out by the same observer using the same instrument. Ideally, several blind and non-successive measurements should be made by, for example, measuring several other children in between. (Under these stringent conditions, and only then, can the SD of a single height measurement be reduced by root n, where n is the number of measurements made.) (4) Studies relating to growth should routinely include the error of measurement of the individual researcher or researchers involved in the study. Where this is omitted, the published data cannot be evaluated. (5) We believe that the 'estimated standard deviation' provides a clear and simple expression of reproducibility, and propose that it should be routinely adopted.

LDV was generously supported by KabiVitrum (UK) and TJW by the Wellcome Trust. We are grateful to Dr J M Walker and Mrs L Vey for useful discussion and help. We are also indebted to Mrs W Couper for preparing the manuscript.

1 Hindmarsh PC, Brook CGD. Auxological and biochemical assessment of short stature. *Acta Paediatr Scand [Suppl]* 1988;**343**:73–5.
2 Tanner JM. Normal growth and techniques of growth assessment. *Clin Endocrinol Metab* 1986;**15**:411–51.
3 Baldwin BT. *Physical growth and progress*. Washington DC: US Bureau Education, 1914. (Bulletin 10.)
4 Krogman GM. A handbook of the measurement and interpretation of height and weight in the growing child. *Monographs of the Society for Research in Child Development* 1950;**13**:3.
5 Albertsson-Wikland K. Growth hormone treatment in short children—short term and long term effects on growth. *Acta Paediatr Scand [Suppl]* 1988;**343**:77–84.
6 Hindmarsh PC, Brook CGD. Effect of growth hormone on short normal children. *Br Med J* 1987;**295**:573–7.
7 Lippe B, Rosenfeld RG, Hintz RL, *et al.* Treatment of Turner's syndrome with recombinant human growth hormone. *Acta Paediatr Scand [Suppl]* 1988;**343**:47–52.
8 Molinari L, Largo RH, Prader A. Analysis of the growth spurt at age seven (mid-growth spurt). *Helv Paediatr Acta* 1980;**35**:325–34.
9 Hoey HMCV, Tanner JM, Cox LA. Clinical growth standards for Irish children. *Acta Paediatr Scand [Suppl]* 1987;**338**:1–31.
10 Caruso-Nicoletti M, Malozowski S, Kibarian M, Barnes KM, Cassorla F, Cutler GB Jr. Short-term variation of growth rate and somatomedin-C levels in normal prepubertal children. *Journal of Paediatric Endocrinology* 1988;**3**:7–13.
11 Cameron N. The methods of auxological anthropometry. In: Falkner F, Tanner JM, eds. *Human growth*. Vol 3. 2nd Ed. New York: Plenum, 1986:3–46.
12 Preece MA. The anthropometric considerations in the evaluation of growth promoting treatments. In: Ranke MB, Bierich JR, eds. *Paediatric endocrinology—past and future*. Munich: MD-Verlag, 1986:22–7.
13 Tanner JM. Physical development. *Br Med Bull* 1986;**42**: 131–8.
14 Rose GA, Holland WW, Crowley EA. a sphygmomanometer for epidemiologists. *Lancet* 1964;**i**:296–300.