DATA ACCURACY AND MODEL RELIABILITY

Jake Chapman

Professor of Energy Systems, Faculty of Technology, Open University.

This paper explores the relationship between the accuracy of the predictions of a model and the accuracy of the underlying model and the data set with which it is provided. Experience with a number of software implementations of BREDEM are described in some detail. Using date from the NHER practical assessment examinations the error rate from trained professional users in estimated - and found to be about 5 errors per 100 data items. The paper describes how careful software design can accommodate this sort of error rate without severe loss of reliability. It also argues that the errors in data input are the limiting factor in the overall reliability of computer models made available to a wide audience.

INTRODUCTION

There is a broad spectrum of models for evaluating the energy performance of dwellings and a corresponding spectrum of users of such models. At one end of the spectrum there are specialist research users who use and develop a model as part of a programme of research into building performance. At the other end of the spectrum is an architect or builder who will occasionally make use of a model as a design tool. Particularly where the design is of a domestic dwelling then the architect or builder will want the process to be uncomplicated and relatively quick and accessible. Indeed if suitable computer models are not available then the designer will usually resort to some simple rules of thumb since energy performance is just one of many aspects of the design. This paper discusses some of the issues of model reliability in the area of simple design tools. Typically the architect will not want to devote more than say an hour to the process of evaluating a number of significant options for the dwelling design.

There is a general belief amongst modellers that the more comprehensive and complete a model is then the more reliable will be its predictions. Particularly as the computational facilities available to modellers have increased so too has the tendency to develop ever more complex representations of the situation. However it is not self obvious that the more comprehensive and complex model will actually result in a better performance - in the sense of giving reliable predictions of building performance.

It is fairly well known that when the same building is analysed by different models a significant variation in predicted energy consumption results: Bloomfield (1). It is also known that when a number of people familiar with a model enter the same building then another large variation in results occurs: Bloomfield (2), Jones (3). The second case is more problematic since with the same building and the same computer model the variation is entirely due to differences in the data input. The differences in data input are due to either differences in interpretation, or measurement error or keyboard error - or some combination of all of these. One of the questions that arises as a result of this second source of variation is the rate at which the reliability of a data set decreases as more data is required from the user.

This can be made more pointed by making two reasonable assumptions. The first assumption is that the rate of error in data input increases linearly with increasing data requirements. Certainly if the predominant errors are of the measurement and keyboard types then this is a reasonable assumption; measuring and entering forty items will generate twice as many errors as entering twenty items.

The second assumption is that the underlying accuracy of a model increases with the logarithm of the number of data items required. By the underlying accuracy of the model I mean the degree to which the model could predict performance given perfect data input. The assumption that this increases with the logarithm of the number of data items indicates a fierce case of diminishing returns. Whether the benefit of additional data is as small as the logarithmic function suggests it will certianly be less than a linear increase. The overall error of the model plus the data can be taken as the sum of these two sources of variation. As shown in Figure 1 with the assumptions outlined above the overall error of the combination will reach a minimum at about the point at which the contributions from each source of variation are equal.

This simple framework could, itself, be complicated further by attempting to take into account that fact that where more data is required then the model's sensitivity to any one data item tends to decrease. But this is a non-uniform effect; some items of data (e.g. exposed wall area) retain the same sensitivity no matter how many other data items are requested. So rather than try to complicate this "model of models" this paper aims to begin the process of quantifying the relationships between the number of data items required by a model and the resulting reliability of the model and data sets as supplied by real users. In particular the paper will focus on the range and types of errors made by users of models in the hope that by better understanding how data errors are made their occurrence can be reduced.

ENERGY AUDITOR AND ENERGY ASSESSOR

Energy Advisory Services Ltd has produced a wide range of micro-computer programs based upon the BREDEM model for predicting the energy performance of dwellings. The first such program, known as Energy Auditor, required the user to measure the areas of all the external elements of a dwelling and enter these and appropriate U-values into the computer. The program also requires information on ventilation factors, heating system, heating controls, water heating, cooking and other appliance use in the dwelling as well as site factors such as orientation, overshading and wind shelter. Once all the data was entered then the program predicted the energy consumption in the dwelling and enabled the operator to assess the installation costs and savings of a wide range of improvements to the dwelling and its equipment. This program was independently tested against a set of field data by Henderson and Shorrock (4) and was found to perform well over a very wide range of circumstances.

Although Energy Auditor performed well in the sense of providing good predictions of running costs it was not suitable as a general auditing or energy labelling process since it took at least two hours to complete the assessment of a dwelling. What is more most people with experience of assessing dwellings could predict which measures would show a short payback time in a dwelling without having to go through the complicated measurement and calculation procedure. EAS Ltd was therefore asked to develop a simpler procedure and program for assessing dwelling energy performance and the likely benefits of insulation and heating improvements. The simplified program, know as Energy Assessor, required the assessor to select a standard floor plan from a menu of alternatives and to then enter two to five dimensions on that floor plan. In addition the surveyor had to enter the number of storeys, the age of the dwelling, the built form, the heating system and its controls, the water heating system, the type of glazing and any additional insulation installed. A survey using this program could be completed in an average time of 20 minutes. The program used the age of the dwelling to deduce an average storey height; this combined with the number of storeys and the floor dimensions enabled the wall area to be calculated. The roof areas was taken as the same as the ground floor area. The window areas were estimated from correlations previously established between dwelling floor area and window area for dwellings of different ages and built form. In short the Energy Assessor program required a great deal less information to establish the same data about the dwelling as Energy Auditor.

The Energy Assessor program was developed as part of an BRE investigation into the viability of energy labelling. The main contractors for that investigation, The Energy Conscious Design partnership, also carried out a side-by-side field trial of Energy Auditor and Energy Assessor. Two expert users used the programs to audit eighteen dwellings in London. This side by side comparison of the two programs came to a number of important conclusions. First some defects in the simplifications made in the Assessor program were highlighted; for example it was found that the default storey height used in Assessor was unreliable and that a substantive improvement in accuracy could be obtained by requesting the storey height as a data item. Second it showed that the general predictions of the two assessments were similar, both in detail and overall. Figure 2 shows some of the comparisons.

Finally it also showed the weakness in the Auditor program. In all the cases where there was a discrepancy between the two programs a check on the data input to the Auditors program revealed that there was an error in the data. In the words of the authors of the report to BRE "During the analysis of the results we found ourselves frequently checking Auditor's results against those of Assessor and not vice versa. This was due to the large number of inputs required for Auditor, which made it far easier to make a mistake during data input while visiting a site": Oreszczyn and Doggart (5). Unfortunately there was no record kept of the frequency of data errors - as in most cases the researchers simply corrected the errors and were pleased to spot them before someone else did.

THE NHER EXAMINATIONS

The National Home Energy Rating scheme has been described in an earlier paper in this conference. In order to qualify as NHER assessors people have to attend an NHER Training course and then pass examinations which include the practical assessment of dwellings using the NHER software. The main function of the Training Courses is to establish a set of conventions and procedures to be used in the assessment of dwellings and to practice using these and entering data into the computer. The people attending these courses and sitting the examinations are building professionals; mostly architects, heating engineers, and building surveyors with a few academics from schools of architecture and building science. There is a significant variation in previous computer experience amongst the trainees but this has not been reflected in the examination results.

The NHER scheme has different procedures and computer software for assessing existing dwellings and new dwellings. Existing dwellings are assessed using a program which is conceptually derived from the Energy Assessor program referred to in section 2. The program minimises the number of dimensional data items required both the reduce the time of the field survey and to maintain a high level of data accuracy. New dwellings are assessed using a program which is conceptually similar to Energy Auditor in that areas are extracted from plans and entered into the computer along with other aspects of the dwelling specification (U-values, heating systems etc.). The practical examination for New Build assessors requires them to complete a number of assessments from plans. The practical component for the field audit examination requires the assessor to assess an existing dwelling. In both cases there are large numbers of professional people analysing the same dwellings using the same computer software in a context in which their data entries into the computer can be scrutinised and subsequently evaluated. Indeed the way that the examinations are marked is that the data that the trainee enters is compared to the correct set of data and marks deducted for each significant error. The examiners pay little attention to the overall energy prediction since a correct prediction can (and does surprisingly often) arise as a result of twomutually cancelling errors.

At this point it is worth emphasising a point that may come as a surprise to anyone who has not been involved with this sort of detailed comparison of data entry into a computer program. The fact is that no one gets it completely right. This came as a shock to myself since as designer of the software, the conventions, the training courses and the examinations I expected my own assessments to be "correct" i.e. perfect. Reluctantly I have to admit that they never are. Each time I compare my assessment to those of a group fo 10 or 20 examinees then I will uncover several data items which I have got wrong. Until one has had this experience there is a tendency to attribute errors to carelessness or incompetence. However when confronted with one's own fallibility it becomes clear that when analysing something as complex as a building there are an enormous number of factors and conventions that have to be remembered and applied correctly - and that it is inevitable that some will be forgotten or overlooked.

CLASSIFICATION OF ERRORS

In general it is easy enough to identify an erroneous data entry; however it is often very difficult to understand the source of the error. The difficulty often lies in the fact that the error is due to an incorrect perception or understanding. It is hard to "see" a misconception or to shift one's own perception. However detailed checking of the practical assessments for both the New Build and Field Audit NHER examinations has lead to the identification fo five different categories of errors.

1. Observational errors

These errors are where the trainee has simply failed to notice some detail or aspect of the dwelling being assessed. This is a common source of error in the field audit situation since there are a large number of items to be checked in the assessment. In the new build situation observational errors are rarer - though in some cases people have not noticed that the dwelling being assessed was a mid-terrace version of a house type (see later example)

2. Conceptual/Mapping errors
These are often the most difficult errors to identify since they result from the trainee having an incorrect mapping between what is observed or specified in the real world and what has to be entered into the computer model. This sort of error would almost never arise in a research context or where the main users of the model had a very high level of expertise in the modelling process. However it is relatively common when a computer model is used sporadically by professionals who use different types of models of dwellings.

3. Convention errors
As mentioned earlier it is essential to have conventions for arbitrating in cases which are ambiguous. Within the NHER system there are conventions for deciding how measurements should be taken (e.g. window dimensions are the size of the hole in the wall since the program has built in frame factors), for deciding what is a conservatory and what isn't, for what constitutes a draught lobby, how to delineate the two zones of the dwelling and so on. No matter how carefully these conventions are elucidated there will always be borderline or ambiguous cases - and trainees will forget to apply certain conventions. So this is a common source of error. However the situation is far better than without any conventions. Initially the MKECI did not specify measurement conventions and there were large variations in the evaluations.

4. Measurement errors
Measurement errors are self explanatory. They are also sometimes very surprising. By careful measurement of plans we have discovered that architectural drawings are not usually accurate to better than 3%. There is often a conflict between the plans and elevations. Parallel walls are not parallel so that the width of a building may be different at one end to the other. Within the examination process some tolerance is permitted before a measurement different from the examiner's is counted as an error.

5. Keyboard errors
Both the Builder and Home Rater programs require about 125 data items in order to characterise a dwelling. Some of these date items are entered using many keystrokes; for example wall areas are usually entered as ten pairs of dimensions, each dimension involving 4 key strokes. On average it takes about 500 keystrokes to enter the 125 items. Observation of people skilled with keyboards suggests that the minimum error rate is about 1 per 100 keystrokes for this type of work (touch typists and data entry clerks do much better than this - but normal users who use a range of software and who have to access function keys, typewriter keys and numeric keys as well as cursor keys are likely to have error rates above 1 in 100). So the issue is not whether errors are made but whether they are spotted and corrected.

An Example

In order to illustrate these different categories of errors spend a few minutes looking at the floor plans and elevations in Figure 3. These have been taken from an NHER examination. The examination asks the trainees to assess a house which is clearly identified on the site plan as a mid-terrace house.

It is quite common for plans of terraced dwellings to be identical for mid and end terrace versions apart from a window or two. Hence the floor plans that are shown are actually those for an end-of terrace (note the bathroom window on the side wall). Trainees who were not on the look-out for this included the bathroom window in the assessment of house 72 even though it couldn't be there. Some assessors assumed the window would move around to the rear facade of the dwelling - thereby ignoring the evidence in the elevations. These are classed as observational errors.

The better heated zone in a dwelling is referred to as zone 1 and is defined as follows. Imagine all the room doors in the dwelling are closed and all the full height cupboard doors are open. Imagine standing in the lounge (or living room). Everywhere you can now walk without going up or down stairs is part of zone 1. Using this definition the whole of the ground floor of the dwelling is part of zone 1. However there is an ancillary clause which states quite clearly that zone 1 ends at the bottom of the stairs when these enter the living room directly. This is to avoid ambiguity of knowing how far up the stairs to go with zone 1 in such cases (the program takes the increased interzone heat transfer into account by asking about the location of the stairwell directly). Since zone 1 stops at the bottom step of the stairs how much of the front wall of the dwelling is to be classified as wall external to zone 1? Until this issue was clarified half the trainees opted for the whole wall, the other half stopped at the edge of the stairs. This illustrates how difficult it is to tie down conventions of this type - and how easy it is to make a convention error.

Walls which are covered by unheated spaces are sheltered by the unheated space to some degree. The NHER system includes a simple convention for reducing the U-value of sheltered components by a fraction to take account of this shelter effect. In the case shown in Figure 3 the meter cupboard or bin store adjacent to the front door shelters part of the frong wall. (The porch above the front door may also shelter part of the wall if there is a soffit to the porch). Trainee's who missed entering this wall area as having a reduced U-value either made an observational error or a conceptual/mapping error (which would be that the bin store/meter cupboard made no difference).

The floor plan shown in Figure 3 is one of those where the dimensions of the dwelling vary overs its length. Not by much, but wehn multiplied by another number to get the floor area the effect is enought to generate different answers. However the largest source of variation in the floor area was due to measuring to either the inside surface or the centre line of the walls. The NHER convention is that one measures to the inside surface of party walls and to the centre line of external walls. Since we are analysing a mid-terrace case the correct measurement is from inside surface to the inside surface across the plan, and from the centre line to the centre line up and down the plan. Only a few trainees applied this convention rigorously.

The largest source of variation in measurement was with the windows. Those who are used to reading such plans will recognise a number of standard window sizes in use; 1.35 x 1.05 and 0.9 x 1.05. However if you take dimensions of the elevations you will find a range of sizes - and the widths will not be the same as those measured on the plans.

Some people made straight measurement errors and simply got some windows wrong by a factor of two (using the wrong scale on the scale rule?) or the length of the dwelling wrong by 0.5m (a scale reading error). It is also possible that some of the errors attributed to measurement errors were in fact keyboard errors - from the examiner's perspective it is almost possible to distinguish between these sources of error (since they are both random).

Another source of variation with this example was the calculation of the wall U-value. The wall specification was as follows "Facing brick, 50mm cavity, 125mm Turbo block (conductivity 0.11 W/mK), 12mm plasteboard on dabs." The NHER program has a u-value calculator built into it so that all the program user has to do is specify the correct number of layers and the thickness and composition of each layer. The U-Values calculated by the trainees varied between 0.59 and 0.54. There are two sources of variation in this calculation. The first is the thickness of the brick layer - the range was between 102mm and 114mm. The second variation is the thickness of the air gap between the plasterboard and the blockwork. Some ignored it altogether, some put it at 5mm and some put it at 10mm. Any offers for the "right answer". Should the calculation take into account the bridging effect of the mortar between the blocks? and the bridging of any gap between blocks and plasterboard by the dabs? This example makes it clear that for some items the "correct" answer is simply unknowable.

FREQUENCY OF DATA ERRORS

The above discussion of the things that can go wrong in an assessment may have created the impression that the whole process is incredibly error prone. This is not the case. The way that the practical examinations are marked is that each data entry item that is wrong loses the candidate one mark. Where one measurement causes several data entries to be wrong then several marks are lost (since such measurements should be checked with more care!). The examiners do not deduct marks when they identify a convention confusion that has not been clarified by the Training courses (as happened in the above example). Nor do they deduct marks for small variations in dimensional measurements or where the plans and elevations are inconsistent (as in the windows examples above). Everyone starts with 40 marks and the pass level is 30/40. Figure 4 shows the distribution of marks for 100 practical assessments of dwellings from plans. Note that there was one, just one, perfect assessment.

The distribution shows that the median number of errors was 5; and the mean number somewhat larger than this (note that there were four cases where the overall score was less than 20). This seems a reasonable estimate of the error rate to be expected from this type of model application. It might be argued that since these were new trainees their error rate might be higher than the long run average. Against this they knew that they were being examined on these assessments so they would be doing them with more care and thought than might be applied normally.

In the field audit case there is a higher incidence of errors. Again the total number of data items required is about 125 - but more of them are qualitative (such as the presence or absence of flues, low energy lights, additional wall insulation and so on) rather than quatitative. Fewer field audit examinations have been run, so the distribution of results

is more straggly; however as can be seen from Figure 5 the median number of errors is significantly higher at 7. Incidentally in the field audit case the assessors are told to aim to get the overall window area correct to within 2 square metres; provided they do this they are not penalised for incorrect measurements on any one window. In practice no two surveyors measured any one window to be the same size - in some cases the differences were impossible to understand (reading feet instead of metres?). However there were relatively few cases where the total window area was more than 2 sq.m in error.

The field Audit program was designed to minimise the amount of dimensional data that had to be collected from a dwelling. This was done principally to reduce the time required for an assessment. However the results of the Field Audit examinations show that this was also a wise strategy for minimising errors. All items that had to be measured on the field audit examinations were subject to significant variations - usually with no two measurements (in a set of ten) being the same.

SOFTWARE DESIGN STRATEGIES

When the NHER software was designed it was realised that there would be a significant incidence of data entry errors of all types. The data entry routines were therefore carefully designed to minimise the impact of any one error. The main strategies employed were;

A. Use the selection of items from a menu rather than require the entry of a number. Also the menu's were designed so that choosing between adjacent items did not have a significant impact on the rating.

B. Provide clear on-screen help explaining what is required at each data point.

C. Permit areas to be entered as pairs of linear dimensions thus eliminating "side-calculations"

D. In the case of the Field Audit program the correlation results available for predicting floor and window areas from the other data on the dwelling are used as a cross check on the entered values. Where the entered values are significantly different from the estimated values then the user is warned to check the data.

BREDEM models are relatively insensitive to the area and u-value of any one external item; typically a 10% change in the area of an element will lead to a 1% change in the overall energy consumption. However, like most energy assessment procedures, the model is very sensitive to the degree day region and the heating system efficiency which both affect the fuel use fairly directly. The total floor area of the dwelling is significant in the in the NHER computation (since it is based upon running costs per square metre) so this has to be assessed with particular care. The use of menus of heating systems and degree day regions (with fixed values associated with each menu option) reduces this problem slightly - but it is still the area of greatest sensitivity.

The overall target for accuracy of the NHER system was ±10% and theere are indications that this has been achieved. The spread in NHER assessments from the examination examples is largely within a 10% margin as shown in Figures 6 and 7. (Note that these include assessments by some trainees who were subsequently failed). In practice it is extremely doubtful whether U-values and heating system efficiencies are known to this level of accuracy.

Experience with teaching assessors how to use the software and apply the conventions has demonstrated that there are a number of improvements that can be made to both that would significantly improve the reliability of the system. For example in the Field Audit program the assessor has to select a floor plan shape and then enter dimensions on that shape as it appears on the screen. In many cases the shape on the screen can only be related to the sketched floor plan shape on the assessors worksheet by a rotation and inversion. Transferring numbers from the sketch plan into the computer under those circumstances is fraught with potential for error. It is a relatively simple task for the program to permit the assessor to rotate and invert the shape on screen before entering the data. This will significantly enhance the ease of entering this critical set of data.

CONCLUSIONS

Where energy performance models are to be used outside the research environment then the design of the model and associated software must take into account a significant rate of error in entering data into the computer. The results reported in this paper refer to the evaluation of 100 dwellings from plans and thirty dwellings by field audits. The data sets in each case contained about 125 items and the average error rate was about 7 errors per

assessment. Since the software had been designed to accommodate a level of error in the data input this error rate in the data input did not cause more than a 10% error in the overall assessment. Without designing for catching and absorbing errors the error in the assessments would have been significantly greater (estimated to be ±20%).

It is unlikely that computer models in widespread use can ever guarantee sets of data with error rates less than those found here. Indeed without training courses and detailed instructions on how the model works the error rates would be much greater - particularly with respect to conventions and the meaning of certain data items. This also implies that in the trade off between model complexity and data simplicity the designer of general purpose software should consistently move toward reducing the data requirement. An extra algorithm that might improve the model performance by 2% will not realise that improvement if it requires an additional 10 data items to be specified. From the results presented here it appears that the accuracy of the predictions of computer models lies is limited by the error rate of data input and not significantly by the underlying model itself.

REFERENCES

1. Bloomfield, D.P. "Design tool evaluation - bench mark test cases", IEA Task 8: Technical Report T8-B4 1989 (avail from BRE, Watford, UK)

2. Bloomfield, D.P. "The Influence of the user on the results obtained from Thermal Simulation Programs. Proc.5th Intl. Symp. on use of comp. for Env. ENG. Related to buildings. Bath 1986.

3. Jones, L. "Analyst as a factor in the prediction of Energy Consumption" Proc.2nd Intl. CIB Symp. on Energy Conservation. Copenhagen 1979.

4. Henderson, G. and Shorrock, L.D. "BREDEM - The BRE Domestic Energy Model - testing the predicitons of the two zone model", Build.Serv.Eng.Res.Technol 7 (2) 1986 p87.

5. Oreszcyn, T. and Doggart, J.V. "Simplified Auditing and Labelling of Houses", final report to BRE 1987.

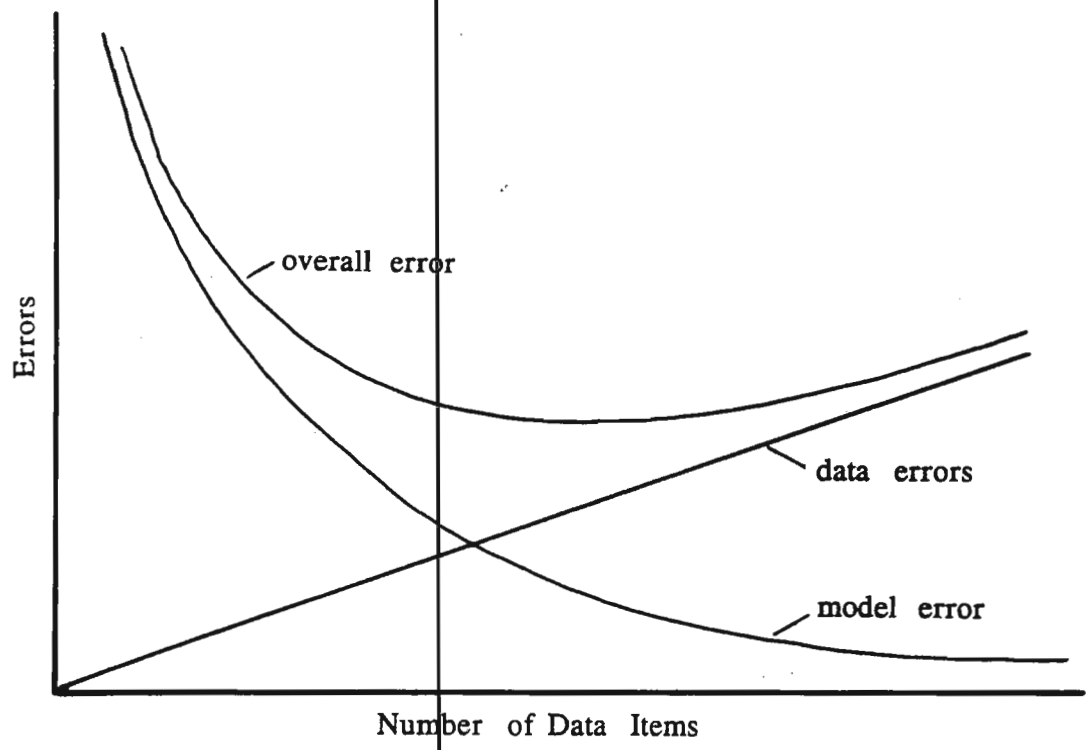


Figure 1 Combining model and data errors

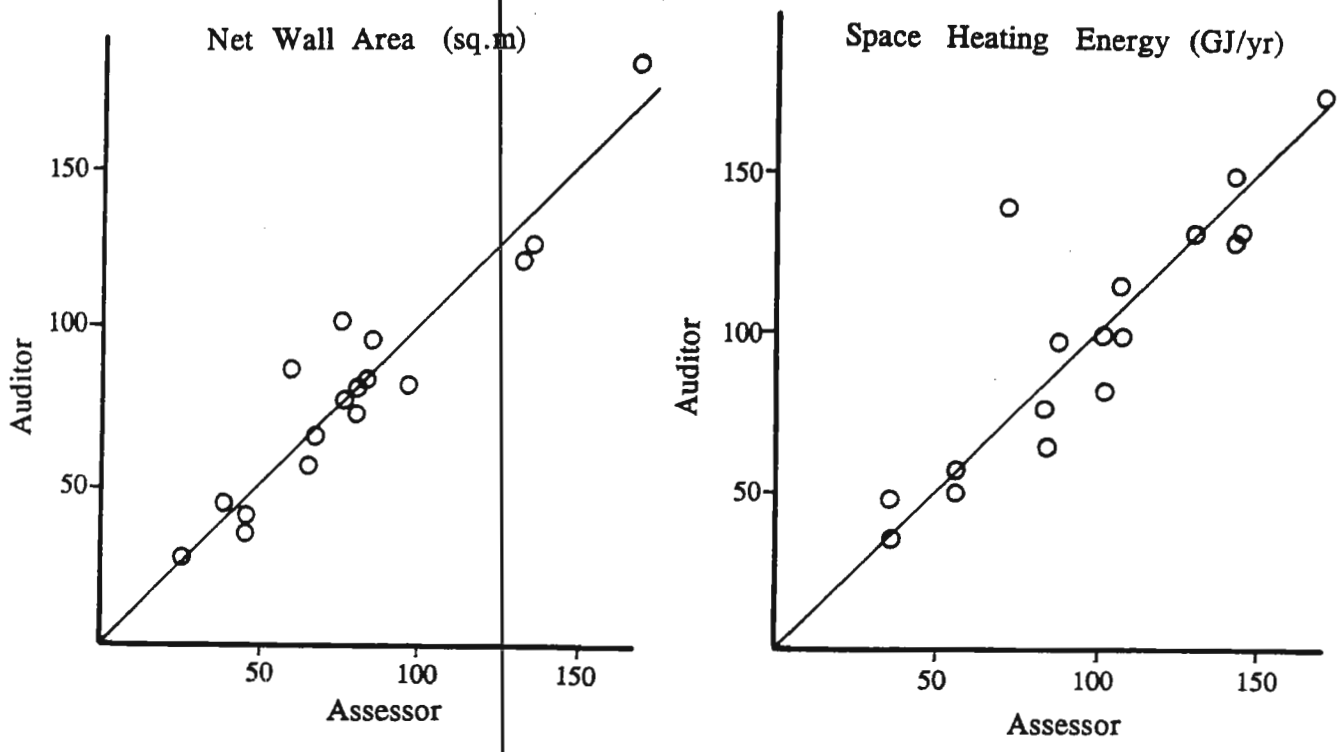
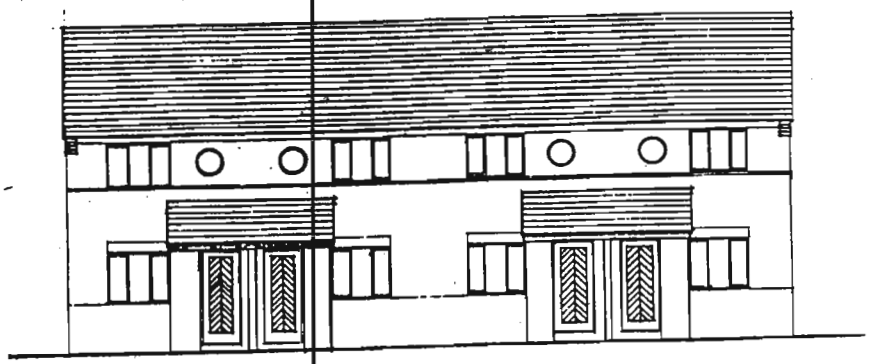


Figure 2 Comparison of results from Energy Auditor and Energy Assess



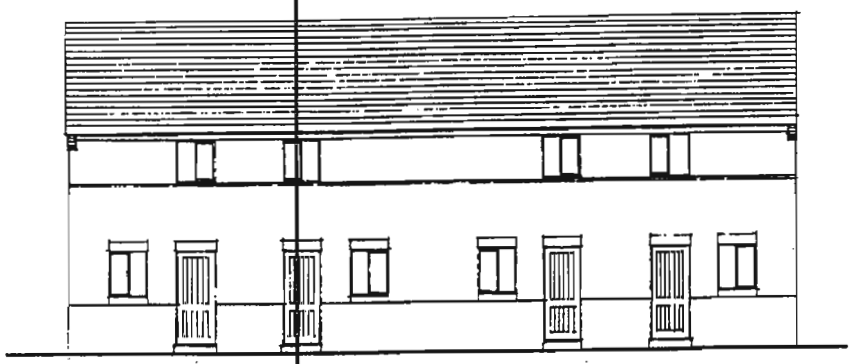


Figure 3(a) Floor plan of NHER New Build Examination example

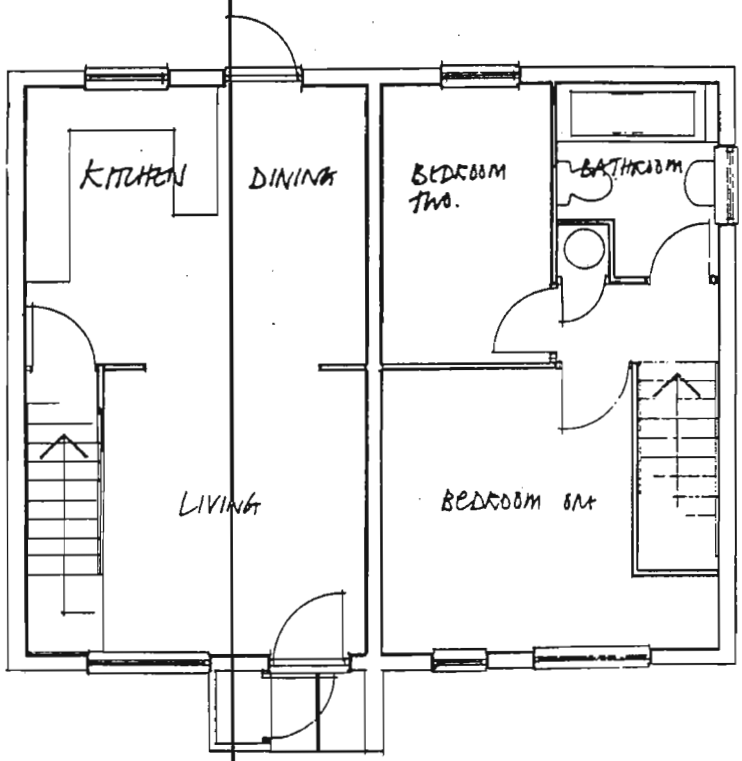


Figure 3(b) Elevations of NHER New Build Examination example

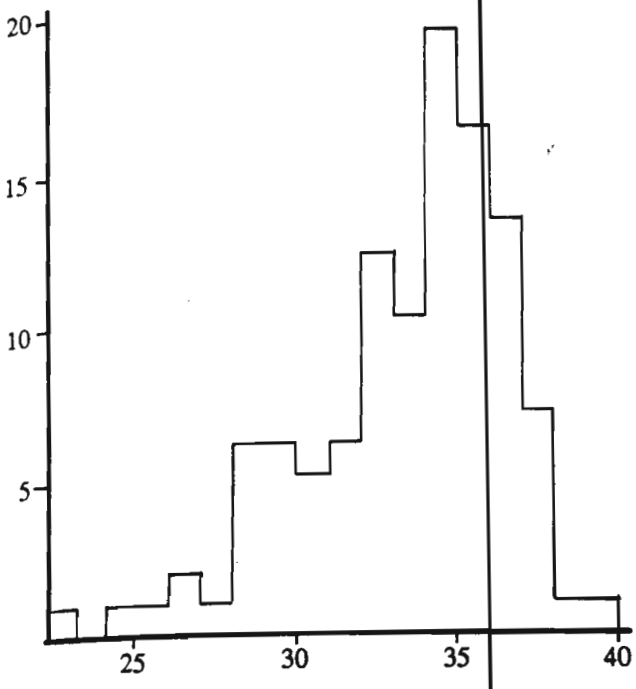


Figure 4 NHER New Build exam scores

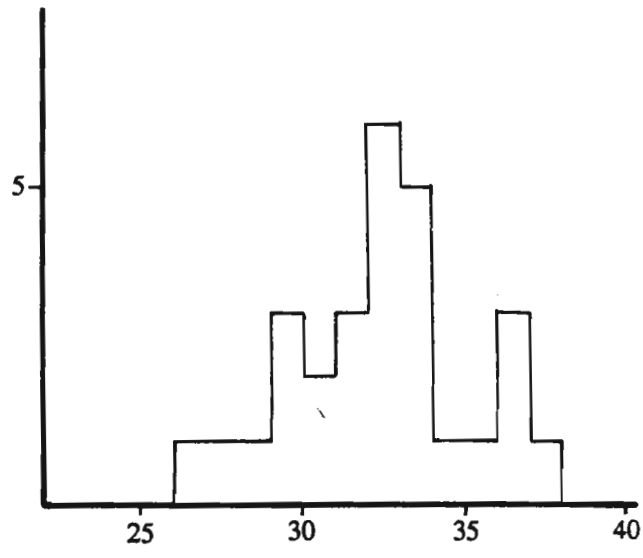


Figure 5 NHER Field Audit exam scores

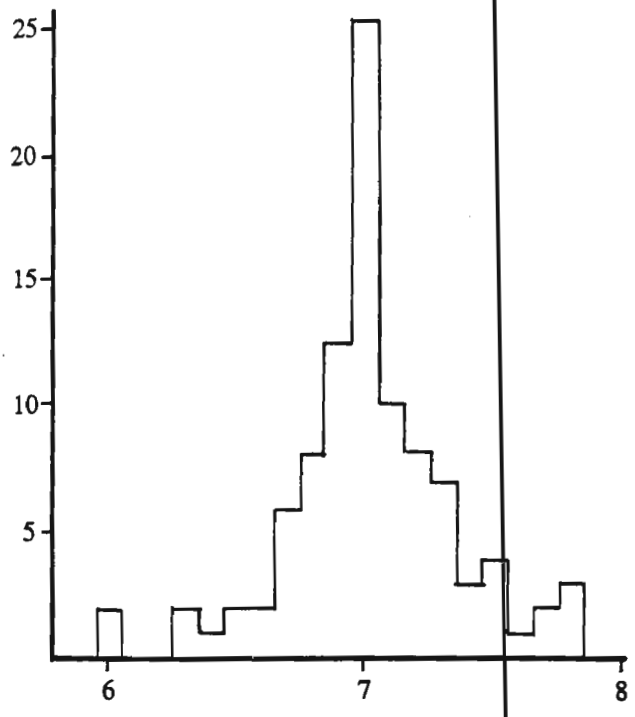


Figure 6 Ratings of New Build exam example

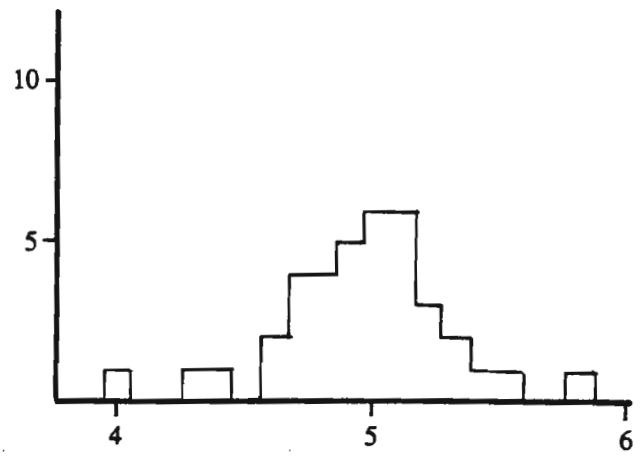


Figure 7 Ratings of Field Audit exam example