Is Job Evaluation Scientific?

Bill Cowan

There are a number of reasons why job evaluation is not scientific, despite the appearance - the measurements, graphs, calculations. The simplest reason is that the aim of job evaluation systems is to say what rewards people should receive. In fact, job evaluation sytems do not specify the actual amounts which people should be paid, but instead specify the order of value in which jobs should be placed, and how much should be paid for each job relative to the others. But this doesn't make any difference. The issue which job evaluation is addressing is still one of "distributive justice" - of who should get what, how the cake should be divided. And science does not provide answers to such questions. Science depends on empirical proof to test its theories. What empirical proof could be found for a theory which states that one person should be paid twice as much as another?

To give a concrete example: the Paterson system of job evaluation says that people should be paid more when their job entails a higher level of decision-making. This is a non-scientific proposition. How could you prove it or disprove it? If one person drives a truck, delivering heavy goods, and another person sits in an office deciding where the goods should be delivered, perhaps the office worker should be paid more because that person is making the decisions. Or perhaps the truck driver should be paid more, since the work is more strenuous? Science cannot provide the answer. These are matters of opinion, matters of negotiation, whose final outcome depends upon the relative power of the parties involved.

Psychological research

People who develop job evaluation systems sometimes claim their ideas are based on psychological research. The main argument derived from such research is that people are

primarily interested in how well they do relative to others. The claim is that people will be satisfied if they get more money than someone else who does less valuable work, or also satisfied if they get less money than someone who, they feel, is doing more valuable work. But these ideas don't take you very far.

For a start, how will people decide what is more valuable and less valuable work? It may be possible to do this in a laboratory experiment with simplified and standardised tasks, but not in the complex real world - we can expect people to disagree (and if they disagree, they can't all be satisfied). Secondly, even if people are interested in how well they are doing relative to others, this doesn't mean this is all they are interested in. They might be interested in having enough money to survive, to buy the bare requirements of life - in other words, in the absolute amount they earn, not just the relative amount. It is quite possible for job evaluation systems to indicate wages below the poverty line. If people were paid at this level, would they be satisfied just because they were earning more than even lower-paid workers? Thirdly, you can't in any case take results from a few laboratory experiments (mainly American) and expect them to describe other people in very different situations, with different values and needs. These problems undermine any claim that job evaluation theory has a foundation in psychological research. In fact, one could regard such claims as an attempt to disguise the fact that job evaluation systems impose their own value judgements, with no scientific backing.

The methods of job evaluation systems

When negotiating about a job evaluation system, it is important to know at what phases in the job evaluation procedures these value judgements come in - for these are the "weak points" in the systems, the areas where (as far as a scientist is concerned) negotiating parties are quite entitled to disagree.

Unfortunately, though, while these are weak points from the point of view of scientific validity, they are not necess-

arily the easiest to negotiate. The non-scientific assumptions may be buried so deeply in any given system that to reject the assumptions may mean rejecting the system as a whole. This may or may not be a union's intention. It is safe to say that job evaluation systems present affected employees with both benefits and disadvantages - and workers and their representatives will want to weigh up the balance in their particular circumstances. Nevertheless, even if a union decides that it does not want to work within a job evaluation system, it is useful to identify where the system is scientifically faulty, so that desired modifications can be argued for, without being stopped by the mistaken reply: "You can't touch it - it's scientific!" We can start by looking at three phases in job evaluation:

1. The first phase is selecting the criteria for job analysis and deciding how these will be applied. (For instance, in the Paterson system, the "level of decision-making" is chosen as the criterion for distinguishing between jobs.)
2. The second phase is the actual measurement process, where different jobs are assessed in terms of the criteria selected in phase 1.
3. the third phase is the grading of jobs on the basis of the measurements taken in phase 2.

We will stop there. There is, of course, a final phase which is most vital both to employers and employees, and that is making the link between the grading of jobs and the grades of wages. This last stage raises problems of a different kind. Having decided (by phase 3) that one job is of higher "value" than another, the decision must then be made: how much more should that job recieve? This would require a separate full discussion and we rather spell out the criteria against which to judge the first three phases.

Reliability and validity

Now the three phases shown above are very familiar to social scientists who are used to making social measurements in their research. The whole procedure is only reckoned to be methodologically acceptable if they are satisfied that the procedure is (a) "reliable" and (b) "valid".

Reliability has to do with the way measurements are made. In order for a measurement process to be reliable, we want to know that different ways of measuring the same thing will yield the same results, and that if different people measure the same thing - using the same or different methods - they will come up with the same results.

Reliability of measurement is basic to any scientific investigation, but it is not, by itself, enough to ensure "validity". Validity requires, amongst other things, that what people are actually measuring is what they say they are measuring. For example, if an intelligence test was constructed which consisted only of mathematical problems, it might produce consistent, reliable results, but it would not be valid as a measure of intelligence, because intelligence is not just mathematical ability.

Now let's see how job evaluation systems measure up to these two criteria of reliability and validity. In doing so, we will need to distinguish between different types of systems. As we will see, some are weak in phase 1 and relatively strong in phases 2 and 3, while others are stronger in phase 1 (the selection of criteria) but collapse in phase 3.

Phase 1: selecting criteria

In this phase, criteria are selected and applied, to indicate the content of a job, relative to other jobs. Some job evaluation systems employ only one primary criterion. The Paterson system is the commonest example of this in South Africa, and in this system the criterion chosen is the "level of decision-making".

Now to use only the "level of decision-making" as an indicator of job content makes this conceptualisation invalid. It lacks even what methodologists refer to as "face validity" - one can see at a glance that whatever different aspects go to make up job content, there must be more aspects than simply decision-making, and some of these will be quite independent of the "level of decision-making" displayed in a particular job, eg. the effort expended in the

work, the danger or discomfort experienced, the level of training required, and so on. Any method which starts off by ignoring such scope for disagreement is certainly not a valid scientific procedure.

An advantage of such a job evaluation system, which looks at only one primary factor in assigning jobs to "bands", is that it is relatively simple and fast to implement. This is particularly an advantage to management, but it could also be an advantage to unions if they expect their members to benefit from the successful implementation of the system.

Some dangers should be pointed out, however:

(a) Once one has accepted "decision-making" as the primary criterion for putting jobs into bands, having made this concession (which has no scientific grounding) one cannot expect to make much ground in arguing about phases 2 and 3, because in the Paterson system these can be relatively sound.
(b) Accepting the "level of decision-making" as the primary criterion could be to the disadvantage of semi-skilled and unskilled workers, who will be crowded together in the lowest categories, with little chance of getting out.
(c) The "level of decision-making" incorporated in particular jobs will reflect in part the structure of control in a firm or organisation. If superiors prevent those below them from taking their own decisions, as will be the case in a strongly hierarchical organisation, then people at the top will be paid more and people at the bottom will be paid less.

Single factor systems, such as Paterson's, fail the test of scientifically valid procedure at the first hurdle, in phase 1. Measuring the "level of decision-making" cannot provide a valid measurement of "job content". What about more complex systems? The Peromnes System, for instance, which is quite popular in South Africa, pays attention to eight separate factors (problem solving, consequence of error of judgement, pressure of work, knowledge, the influence of one job on other jobs, the level of comprehension required by the job, educational qualifications required, and the degree of further training needed to do the job

competently). Some systems are even more complex than this, taking account of 26 or more separate factors in assessing job content.

One can definitely say that by applying a variety of criteria, rather than one single primary criterion like "decision-making", there is more chance of including those aspects of job content which different people, in different jobs, think are important. But the story unfortunately does not end in phase 1, for phases 2 and 3 are still to follow. The problem of validity has in fact just been moved to phase 3, where unscientific decisions have to be made about how to combine all the different aspects thought up in phase 1. Are all the aspects of equal importance, or must some be given more weight than others? We will look at this problem in more detail below.

Phase 2: the measurement process

Once the criteria have been selected, the next phase is to discriminate between different jobs on the basis of these criteria. This is the measurement phase, and here (from the point of view of scientific method) the main thing we want to know is: are the methods of measurement reliable? Will different people come to the same conclusions?

In general, this kind of measurement is likely to be more reliable if (a) a fairly rigid measuring procedure is adhered to; (b) if subjective judgements are kept to a minimum; and (c) if the measurements are not too elaborate. But remember that the reliability of measurement does not guarantee overall validity. For example, a rigid measuring system may improve the reliability of measurement, but it may mean that you are measuring the wrong things most of the time, so that the results lack validity.

We can again look first at the Paterson system, as an example of a relatively simple single factor system. In the Paterson system, the usual procedure is to obtain written job descriptions (which should be approved by the person doing the job, and his/her supervisor) and then to take these to a grading committee, who apply the fairly rigid

guidelines laid down in the Paterson system to determine which "band" of decision-making the job falls into. Because the categories of decision-making are quite broad, and because the level of decision-making is the only factor which needs to be assessed at this stage, there is reduced scope for disagreement. There may be argument about borderline cases, but there is not so much room for disagreement as there would be in a more complex system. One would therefore expect the measurements to be fairly reliable at this stage. But one should remember that this gain in reliability comes partly from ruling out all considerations other than the level of decision-making.

In the next stage in the Paterson system of grading, supervisory grades and further sub-grades are made within each of the bands, and here the judgements seem to become more subjective, with more scope for disagreement - ie. more risk of unreliability. The reason for this is that at this stage the judgements become more complex, taking account of various different factors such as work pressure, variety of tasks, etc.

When workers or worker representatives are on the evaluation committee, this is likely to be an area for negotiation, since it is so "visible": by this stage of the proceedings, the evaluation committee is asking more direct questions like, "Should this group of workers get more than this other group, or should it be the other way round?" There is maybe something a bit deceptive about this area for debate, for it can only concern small changes in grading, as the bands have already been settled. For the same reason, though, the unreliability of decisions at this stage of deciding the sub-grades is less serious from the scientist's point of view, precisely because these decisions don't make such a large difference to the overall picture. Overall, one would expect different teams of job evaluators to come to most of the same conclusions about which jobs were in which bands with relatively minor disagreements about the sub-grades within each band. So one expects the Paterson system measurements to be fairly reliable overall, but inaccurate when it comes to details. (And of course most of the relevant "details" have been

pushed aside by taking decision-making as the single primary criterion.)

It is worth noting that a fairly rigid measuring procedure, such as that in the Paterson system, makes it easier to manage consultations with workers and unions at this stage of job evaluation. It leaves some scope for disagreements and negotiation, but not so much scope to threaten the overall design of the pay structure. Both in the job description stage and in the grading stage workers may be consulted. However there is typically little, if any, consultation about how the terms have been set - why "decision-making" has been adopted as the primary criterion, and how the measuring schedule has been drawn up.

We can turn now to "point-scoring" methods of job evaluation, such as the Peromnes system. The idea here is that jobs will be assessed on a number of aspects, and on each aspect (or "factor", or "dimension") a particular job will be rated as earning so many points. These scores are then combined to give an overall rating for the job. We will see in the next section that arbitrary, non-scientific decisions have to be made in order to combine the scores which a job gets on each different factor. In the meantime, we are mainly interested in whether the measurements, on each factor, are reliable.

There are a number of possible measurement problems in a multi-factor system. First of all, because there are more aspects of a job to be measured than in a single-factor system, and because some of these factors require subjective assessment (for example, in the Peromnes system, the "pressure of work") there is more chance of going wrong than in a simpler system. On the other hand, the errors made in measuring one factor may be averaged out by errors in the opposite direction in measuring other factors. So one can't be sure - which is itself, of course, a problem.

Given that uncertainty, it would seem to be important to be able to check on the results. But multi-factor systems are rather hard to understand and to apply, which means that it is more difficult for job-holders, or other people

who are not experts, to keep a check on how the results are being produced. The three main doubtful areas are (a) the reliability of the job descriptions on which the measurements are based, (b) the way in which the scale of points for each factor is arrived at, and (c) the extent to which subjective judgements are made in deciding a job's position on the scale. We will look at these problems, taking the Peromnes system as an example.

(a) The Peromnes System uses verbal job descriptions rather than written ones. Because of the complexity of the system, the person who describes the job may not personally know what to describe, but must rely on the expert valuator to ask for all the necessary information. One can argue that this can lead to fairer evaluations, since there is less opportunity for the job-describer to emphasize or exaggerate features of the job to selfish advantage. But it could equally lead to a failure to provide all the relevant information, for the same reason: the job-holder doesn't understand the significance of the questions which the valuator is asking. Also, because there is no written job description and because it requires an expert to evaluate the verbal job description, it is more difficult to carry out an independent check on the valuator's measurements.

(b) The problem of scaling can be illustrated by an imaginary example. Suppose one factor of a job is the amount of formal education which the job-holder needs to possess. If one job requires a standard ten education while another requires a standard eight education, the first job might earn ten points more on the education factor than the second job. But another job requires a university degree. Should it get an additional ten points? Or an additional twenty points? Or thirty, or two? There is no scientific answer to this question, because there is no way of knowing that a standard ten education is "twice as much" education as standard five, or that it is twice as valuable or four times as valuable, or whatever. We can agree that more education is more valuable than less education, but we can't say how much more, and we can't say that the value increases steadily as

you go up the scale, or anything like that. Despite this fact, most multi-factor job evaluation systems arbitrarily lay down how many points can be awarded for different levels of education, different levels of stress, and so on. This is not scientific measurement, but a way of dressing up opinions as numbers, and then applying these opinions as a routine procedure. The opinions are embedded into the measurement process when the system is designed.

(c) The complexity of multi-factor systems requires that a greater number of subjective interpretations have to be made before a job is given a final score. For instance, in the Peromnes system, one of the factors to be considered is how much time is necessary to achieve a level of competence in a job. It requires subjective judgement to decide when a job-holder is just competent, rather than super-competent or a little incompetent. And if it takes one person six months to become competent it may take another only three months to reach this level - so again judgements have to be made about what is "normal". Obviously, the more subjective judgements have to be made, the more chance there is for unreliability, and as stressed above, such subjective judgements are even more uncertain if it is difficult to check them against others' opinions, as is the case when a job evaluation is so complex that only an expert can operate it.

In summary, a scientific enquirer would treat the relative scores awarded to jobs on the different factors in a multi-factor system with a great deal of scepticism. There is no good reason to suppose that the figures are accurate. Even so, this is probably not the major problem with multi-factor job evaluation systems. The great unsolvable problem comes when you try to combine the separate scores awarded for different aspects of a job, in order to come up with a single overall score for the job.

Phase 3: grading

Having selected the criteria for distinguishing between

jobs in phase 1, and having made measurements according to these criteria in phase 2, the task in phase 3 is to convert these measurements into a means for ranking jobs. In the case of a single-factor system such as Paterson there is no immediate problem at this stage. Once jobs have been catergorised according to their "level of decision-making" (phases 1 and 2), this automatically places them in one of the six bands. The ranking which the different jobs receive on "decision-making" simply becomes the band ranking of the jobs. No one can object at this stage. The time for objections was earlier: what grounds were there for choosing "decision-making" as the criterion of job content? The grounds were inadequate, and for that reason the Paterson system lacks phase 1 validity. If this is overlooked, then the Paterson system cannot really be attacked for failure in phases 2 and 3, except in points of detail. (Of course, the other side of this coin is that, because of its phase 1 failure, no amount of refinement in phases 2 and 3 could render it into a scientific system.)

What are the "points of detail" in the Paterson system which can lead to difficulties in phases 2 and 3? First there is the problem of borderline cases, where a grading committee disagrees about which decision-making band a particular job should occupy. This is a problem of measurement (phase 2). But the more important "details", as far as employees and their representatives are concerned, are probably to do with the sub-grades which are allocated to jobs within a "band".

After a job is fixed in a band, there is still the question of whether it is a supervisory job or a non-supervisory job, since each band (except Band A) is divided into a supervisory and non-supervisory grade. Usually there would be little difficulty in deciding, because the structure of control in most large organisations is fairly clearly defined. So there is no major measurement difficulty here, deciding which jobs are supervisory and which are not - but there is a problem in deciding what difference this should make in terms of basic pay. The Paterson system has already made this decision. Supervisory jobs should receive more basic pay than a non-supervisory job in the same band, but

not as much as a non-supervisory job in the next band. Thus the criterion of supervision versus non-supervision always has less effect than the primary criterion, the "level of decision-making". Is this right or wrong? That is surely a matter of opinion.

When one comes to sub-grades it becomes even more a matter of opinion. No clear guidelines are laid down for how different jobs should be allocated to sub-grades, but it seems that a variety of factors can be considered. Le Roux (in this edition of SALB) suggests that these factors can include the variety of tasks in a job, the "length of cycle", the pressure of work and "tolerance and precision". If this is so, which factor is seen as most important, and why? And why should all these factors be less important than the "level of decision-making"? For example, even if a job in Band B is exceptionally varied, non-repetitive, precise and high pressure, it cannot escape to Band C, unless its "level of decision-making" fits into the Band C category.

As suggested earlier, sub-grading can provide an area for negotiation, precisely because the rules for sub-grading are not clear. But this should not obscure two facts. Firstly, the area for negotiation has already been decided, because the additional factors now being considered cannot move a job out of its "decision-making" band. Secondly, when the additional factors are being considered, there is no scientifically valid way of deciding whether one factor, such as work pressure, should carry more weight than another factor, such as the "variety of tasks".

This difficulty in knowing how to play off different factors against one another only hits the Paterson system at the level of sub-grading. But it is a problem which is right at the heart of multi-factor systems.

Multi-factor systems, which look at more aspects of job content than simply "decision-making" have a better chance of preserving phase 1 validity. They may be complex and generally less reliable in the measuring phase, but it is in the third phase that they really knock down the hurdle and fall to the ground in a crunch of non-scientific calc-

ulations. The problem is this: that having taken measurements of different aspects of a job, you now have to combine these figures to come up with a single score.

The problem is easy to illustrate. Suppose that two of the aspects considered relevant are (1) physical effort, and (2) mental effort. And suppose two jobs, Job A and Job B, have been assessed on these two criteria, and have been given scores as follows:

	Physical	Mental	Total
Job A	10	5	15
Job B	5	10	15

Now if the two aspects, mental and physical effort, are held to be equally important, then (as one sees from the straightforward totals) the two jobs must be ranked on a par. Each job collects a total of 15 points. But suppose physical effort is counted as being twice as important as mental effort, or vice versa. Then the scores on these different aspects must be weighted differently. If physical effort counts for twice as much as mental effort, then five points on physical effort is equivalent to ten points on mental effort. So Job A must be graded higher than Job B. But if mental effort counts for more than physical effort then Job B will end up with a higher score than Job A. In other word, however accurate the scores are on each factor, the final result is still indeterminate, because there is no scientific way of deciding how to weight the two factors.

In practice, there are two ways in which different factors can be weighted in a multi-factor system. If the same measuring range is used for each factor - for instance, "level of education" could earn up to a hundred points, "level of stress" could earn up to a hundred points, and so on - then the weighting is done after the measurement. You take each score and multiply it by a number, a different number for each factor, which ensures that the points earned on a factor which is regarded as important count for more than the points earned on a factor which is regarded as unimportant.

(These are of course matters of judgement, not matters of science.) The other way of weighting different factors is to give them different ranges of measurement, so that "level of education", for instance, might be able to earn up to a hundred points while "stress level" could only earn up to a maximum of twenty points. Either way, the effects are the same, and express the system designer's own view of which factors are more important than others.

Following this road produces a series of measurements for each job - a score or a ranking for each of the factors being considered. These scores, however, are not all on one dimension but are along dimensions which may be independent of one another, and so there is no scientific way of adding up these scores to produce a final uni-dimensional ranking. The only condition under which the separate scores or ranks can validly be added up is if the separate measurements are all along one underlying dimension, and even then we would add certain other mathematical requirements. But if these requirements are met, then we can't have phase 1 validity. So job evaluation is caught in a trap. Either it starts off with invalid assumptions, or it ends up with an invalid procedure for combining scores on different factors.

This does not really come as a surprise, because we saw at the outset that the aim of job evaluation is to make value judgements. These value judgements may appear to be hidden, by the trappings of scientificity and a confusion of numbers, but like a disappearing scorpion, they will turn up again if you lift up all the stones. Either the non-scientific value-judgements are made at the begining, in choosing the criteria, or they come back at the end, in making unscientific decisions about how to weight different aspects of a job.