Measurement Of Physical Output At The Job Level

Research and Technical Report 10
Industrial Relations Center
University of Minnesota

WM. C. BROWN COMPANY
Publishers
DUBUQUE, IOWA

Industrial Relations Center Staff

Dale Yoder, Professor of Economics and Industrial Relations, Director
Donald G. Peterson, Professor of Psychology
Herbert G. Heneman, Jr., Research Associate, Assistant Director
C. Harold Stone, Research Associate (on military leave)

Research Scientists

Harland Fox
Theodore Lindbom
Lenore P. Nelson
Josephine S. Welch

Research Assistants

John Boulger
Marian Deininger
LaVerne Hamel
Wayne Kirchner
Thomas Mahoney
Harvey Tschirgi

Administrative and Clerical

Roberta Nelson
Nora E. Eastman
June A. Cheit
Dina Fischbein
Janyce Gervais
Marceline Kusserow
Audrey Boyum
Alyce Pearson
Jeanette Richardson
Flora M. Roth

Faculty Associates

Maurice C. Benewitz
Earl F. Cheit
John G. Turnbull
Lloyd Ulman

Industrial Relations Center Faculty Committee

Richard L. Kozelka, Dean, School of Business Administration, Chairman
F. Stuart Chapin, Professor and Chairman of Sociology
Austin O. Dowell, Professor of Agricultural Economics
George Filipetti, Professor of Economics and Business Administration
William B. Lockhart, Professor of Law
Donald G. Paterson, Professor of Psychology
Lloyd M. Short, Professor of Political Science
Lee S. Whitson, Professor of Mechanical Engineering
Dale Yoder, Professor of Economics and Industrial Relations

Information about the purposes and program of the Industrial Relations Center is available upon request.

Measurement Of Physical Output At The Job Level

by

Einar Hardin

(Research and Technical Report 10)

Industrial Relations Center

University of Minnesota

August, 1951

Published by

Wm. C. Brown Company

Dubuque, Iowa

Measurement of Physical Output at the Job Level[1]

Einar Hardin

Since the beginning of the 1940's interest in the problems of "productivity" has been at an all-time high. A great deal of hesitation and controversy regarding both the concept of "productivity" and the appropriateness of various methods of measurement is evident.[2] It may be sufficient to state here that research into the conditions of productivity may be viewed as attempts to find the conditions under which a certain output comes forth. Output thus becomes the variable which productivity research attempts to predict.

There are several concepts of output, such as total national output of producers' goods or capital formation, output of selected manufacturing industries, output of a given plant, output of a given group of factory employees performing essentially the same operations, or output of a given employee in a given operation. Output may, furthermore, be measured either in value terms or in physical terms.

In this article the discussion of output measurement is confined to the measurement of only one type of output, that of physical output at the job level. First, the term is explained. Second, the usefulness of measurements of that type of output is described. The problems and methods of measuring physical output at the job level are then discussed. This third and main division consists of five sections. Problems of defining output in terms sufficiently accurate for research are discussed in the first section. This section is followed by two sections dealing with problems of direct and indirect measuring methods and particularly problems of validity and reliability. Next is a section on problems of sampling time periods, which is followed by one short section on sampling of employees in measurement of output. The last division contains a suggested contact procedure for outside research workers.

Meaning of "Physical Output at the Job Level"

Physical output at the job level is here defined as the flow of goods through a specified set of operations which are all performed by one and the same operator. The volume of this flow is then defined as the quantity of goods flowing from this set of operations during a specified period of time. To give some examples; the physical output of a power-sewing machine operator engaged in stitching rayon ribbons and company labels on blankets of a given size and type is defined as the number of blankets stitched during a given period of time, such as fifteen minutes, an hour, an eight-hour day, or a forty-hour week. For a butter wrapper the volume of physical output could be in terms of the number of quarter-pound blocks of butter wrapped during, say, an eight-hour day. The physical output of an IBM punch operator may be defined as the

[1] This article was prepared by the writer as a member of the research staff of the Industrial Relations Center at the University of Minnesota. Invaluable guidance and help was given by several members of the staff, particularly Professor Dale Yoder, Professor Donald G. Paterson, and Dr. Herbert G. Heneman, Jr., both in planning the underlying output studies and in editing the manuscript. Much help in the final editing process was also given by Mrs. Helen B. Aaberg, Minneapolis.

[2] See for example Summary of Proceedings of Conference on Productivity, October 28-29, 1946. U.S. Department of Labor, Bureau of Labor Statistics Bulletin No.913, Government Printing Office, 1947, 52 pp.; and Measuring Labor's Productivity, National Industrial Conference Board, Studies in Business Policy No. 15, New York, 1946. 20 pp.

number of essentially equivalent cards punched during, say, an hour, and so forth. The length of the time period chosen depends on the problem at hand.[3]

The above definition may appear to relate only to individual physical output and to disregard group output; usually, however, only a simple aggregation is needed to obtain output when needed.

Usefulness of Measures of Physical Output at the Job Level

Measures of physical output at the job level already serve a number of purposes and may in the future serve additional purposes. Labor cost control permits detection of sources of high labor cost more accurately when the cost of particular operations can be measured than when it cannot; to measure this cost one obviously needs to know the physical output in those operations. The need for such output data is equally obvious when piece-rate or incentive earnings are to be computed. Checks on new production practices like changes in the kind of supplies, tools, and equipment require the same type of data.

Many industrial relations practices are introduced in order to raise "labor productivity." Among these practices are recruitment, selection, training, motion study, and supervision activities. Insofar as these practices aim at increasing "labor productivity" they should be--and sometimes are--checked, which requires measurements of output for the period of time when the effects of such practices are expected to appear in output changes. Present concern with employee attitudes and with social groupings in a factory are based partly on the assumption that they affect output; output measurements are needed in testing this assumption.

Further development of economic theory of employment also seems to require measures of output at the job level. While some economic phenomena may be measured at the national level through national-income analysis, they can probably be explained only at the levels where the changes take place. As Hiram Davis puts it, "After all, actual changes in efficiency or different combinations of labor, capital, and materials are made at the factory, mine, or farm level. Thus it is to this level that we must go, or as near it as practicable, if we want to discover what conditions have really been associated with changes in productivity."[4] It appears necessary, however, to go below the industrial level or factory, farm or mine levels that Davis recommends, and there is considerable reason to believe that the job level will have to be reached.[5] In the theory of the

---

[3]Under certain conditions it may be preferable to measure the time taken for a defined and designated quantity of goods to pass through the set of operations and to obtain the desired value through inversion.

[4]Davis, Hiram S., The Industrial Study of Economic Progresses, Philadelphia: University of Pennsylvania Press, 1947, p. 14.

[5]Changes in the output of a plant may come from many different sources. One set of such sources is connected with the labor force of the plant: manpower with more suitable or less suitable abilities may be recruited, a training program may be instituted or its quality may change, the morale of the labor force may change and this may affect the efforts of the employees. Another set of sources is connected with the physical equipment that is being utilized; better machines may be installed or old machines may run down, and plant layout and transportation methods may change. A third set of sources is connected with the raw materials that are utilized: Brazilian cotton with short fibres may be substituted for Egyptian cotton with long fibres in the spinning mill (or vice versa), or one aluminum alloy may be substituted for some other aluminum alloy. When more than one of the above sources may have yielded output effects, changes in plant output could contain the combined--or confounded--effects of different factors. Measurement of output effects at the job level may then permit isolation of the contributions of factors connected with the labor force of the plant. When the study is concerned with these factors, measurements of output effects at the job level are sought for directly. Given some suitable method for sampling departments and plants, one may subsequently be able to develop an industry-wide or perhaps nation-wide index of changes in "labor's

(continued on p. 3)

firm, questions arise which cannot be answered without measurements of physical output at the job level. Measurements of physical output at the job level are required for an affirmative answer to the question whether the marginal productivity theory can be used for prediction of employment in a firm. Such measurements are also needed in answering the question whether high-wage firms tend to get a better labor supply than low-wage firms.

Methods and Problems of Measurement

Problems of definition

Before measurements of output can be made the variable to be measured must be defined. The following questions may help in giving increased precision to the definition that has already been given and in pointing up three problems of definition. The first question is, what kinds of goods represent the same type of output? In a blanket-sewing department an operator may work on blankets of different sizes, weights, and ease of handling. In a butter-wrapping department an operator may wrap quarter-pound blocks of butter, half-pound blocks of butter, and pound-blocks of butter. In a clothing factory an operator may sew shirt collars of varying sizes and models, and so forth. The question then arises whether and when these various sizes and types of blankets or butter blocks or collars can be lumped together in the definition, or whether they must be defined as different kinds of goods. The answer partly depends on the purpose for which the measurements will be used. If one wants to compare output of one size and type of blanket with output of some other size or type of blanket, the various units obviously should not be lumped together. If one wants to compute piece-rate earnings and if the piece rate is the same for several sizes and types, these can obviously be lumped together. But in most cases the answer is not so clear.

As an illustration, take the following hypothetical study. The researcher wants to find out how conflicts with co-workers affect output, and a satisfactory method for measuring conflicts has been installed. Each operator works on several sizes and types of blankets during the period under observation. Is it then possible to lump blankets of varying sizes and types together? In other words, can physical output be defined as the number of blankets completed, regardless of size and type, or is it necessary to use several definitions, one for each combination of size and

---

productivity." Also, when the study is concerned with changes in internal organization, measurements of output effects at the job level appear to facilitate estimation of output effects of changes in internal organization by eliminating output effects of factors connected with the characteristics of the labor force of the plant. The output effects of changes in internal organization would, in other words, be obtained as a residual. The interest of most economists in studies of physical output at the job level has, until recently, been quite limited, if not entirely absent. Main interest has centered around studies of output and productivity at the industrial or national levels. See for example Fabricant, Solomon. "Problems in the Measurement of the Physical Volume of Output, by Industries, Journal of the American Statistical Association, Vol. 33, No. 203, September, 1938, pp. 564-570; Garfield, F. R., "Measurement of Industrial Production Since 1939" Journal of the American Statistical Association, Vol. 39, No. 228, December 1944, pp. 439-454; Magdoff, Harry, Siegel, Irving H. and Davis, Milton B. Production, Employment, and Productivity in 59 Manufacturing Industries. 1919-1936. Work Projects Administration, National Research Project, Report S-1, Washington, D.C., 1939; Rostas, L. "Productivity of Labour in British, American, and German Agriculture," London and Cambridge Economic Service, Vol. 24, Bulletin No. 3, July, 1946, pp. 78-81; and Stigler, George J. Trends in Output and Employment. National Bureau of Economic Research, 25th Anniversary Series No. 4, 1947, 67 pp. See also Major Sources of Productivity Information, U.S. Department of Labor, Bureau of Labor Statistics, Washington, D.C., June, 1949, 48 pp., mimeographed; and Uses of Productivity Data in American Manufacturing Establishments. U.S. Department of Labor, Bureau of Labor Statistics, Washington, D. C., July, 1949, 17 pp., mimeographed. It is outside the scope of this article to discuss the appropriateness of concepts of output and productivity employed in the above studies. For references to discussions of this question, see footnote 1.

type? Is it meaningful to construct a definition of a "standard blanket" and how can the various blankets be reduced to units of a "standard blanket?" If operators' efforts are different for different types of blankets and if the time of occurrence of conflicts is in some fashion associated with the time of occurrence of certain types of blankets, the output effects of blanket differences and the hypothesized output effects of conflicts will then be confounded in the output data. Hence, it is necessary to eliminate differences due to blanket types. The use of the "standard blanket" construct may perhaps provide such elimination.

The "standard blanket" construct underlies the use of piece-rate earnings or of total standard time as measures of physical output. These two methods may be acceptable if the time standards are set accurately. But how can one check the accuracy of time standards? Repetition of stop-watch studies is expensive and, more important, provides only a partial check: after the time actually used up in each operation or motion is measured, time allowances have to be added, and correction for the estimated skill and effort of the operators that have been timed, the so-called leveling process, must be made.[6] Time study men frequently have a wide latitude in deciding the magnitude of certain time allowances, and particularly when a 66% fumbling factor" is permitted, the accuracy of the time standards becomes dubious.[7] Someone might suggest that time standards are accurate when an operator makes the same hourly piece-rate earnings on work with different sizes and types of goods, such as blankets. The writer doubts that this can be used as a criterion of accuracy. If employees attempt to make a certain amount of money per hour regardless of the type of work performed, this check may easily become ineffectual. The use of differential piece rates, i.e., one piece rate applies for outputs up to a certain limit and another and higher piece rate applies to output above that limit, makes it inadvisable to define output in terms of piece-rate earnings whenever the given limit is exceeded. In case piece rates have been changed by different percentages, measurements of output, obtained by measuring piece-rate earnings before and after the change in rates, are comparable only if the composition of output was the same before and after that change. Such equality is very unlikely to occur.

If accurate measurements of physical output are required, it is probably necessary--as the preceding discussion indicates--to treat each size and type of blanket, to take an example, as a different product. This applies particularly when all outputs in the selected total population of time periods are measured and when the data are to be used for comparison of operators. When methods of random sampling are utilized for estimation of output in the chosen total population of time periods and when main concern is with the output of an unchanged group of operators, it may be satisfactory to define output in terms of piece-rate earnings or total standard time. Further exploration is necessary before it can be known when inaccuracy introduced by the use of piece-rate earnings or total standard time is within acceptable limits.

When time standards have not been established or when the researcher cannot make his own time studies, each size-type combination must be considered as a separate product and the flow of each combination must be measured separately.

The second question that may help point out problems of definition and thus promote increased precision in definition is--How long should the time period be for which the

[6]For one description of recommended practice, see Barnes, Ralph M., Motion and Time Study, New York, John Wiley and Sons, Inc., 3rd Edition, 1949, especially chapters 20,21, and 22.

[7]The use of such a "factor" is not recommended by time and motion study teachers but is actually not uncommon. For discussions of problems involved in setting time standards see Gomberg, William. A Trade-Union Analysis of Time Study. Chicago, Science Research Associates, 1948, 243 pp. and Littauer, Sebastian B. and Abruzzi, Adam, "Experimental Criteria for Evaluating Workers and Operations," Industrial and Labor Relations Review, Vol. 2, No. 4, July, 1949, pp. 502-526.

flow measure is desired? When is a fifteen-minute period appropriate? When is an eight-hour day appropriate? When is a two-thousand-hour year appropriate? The answer depends mainly on the problem that the study attempts to solve. Rothe has advanced an hypothesis that incentives are ineffective when the ratio of the range of intra-individual differences in (fifteen-minute) output rates exceeds the ratio of the range of inter-individual differences in output rates.[8] Testing of this hypothesis requires measurements of fifteen-minute output rates, and the time period thus becomes fifteen minutes. In other cases it may be necessary to know weekly output, and in other cases again, such as in evaluation of testing programs, the researcher may be concerned with yearly output.[9]

The third question that may promote increased precision in definition is whether output should be defined to include only the number of blankets, quarter-pound blocks of butter, or IBM cards that has been completed during the defined period. If the answer is yes, output does not show continuous variation, in the above examples, and the term "counting" may then be more appropriate. This is in itself no important disadvantage. A great disadvantage may however appear when one attempts to relate variations in output to whatever factors are suspected to cause these variations. If an operator completes, on the average, twenty blankets per hour and if one would want to study the relationship between, for example, fifteen-minute oxygen consumption and fifteen-minute output, the observed relationship may be lowered considerably by the very definition of output. This is so because no account is taken of the amount of work already done on a blanket that was not completed at the beginning of the period or that was already completed at the end of the same period. Similarly, no account is taken of the amount of work done during the period on a blanket that was not completed until in the beginning of the following fifteen-minute period.

The importance of this disadvantage depends on the proportion of the given time interval which is, on the average, taken to process one unit of output. If this proportion is only one per cent, or, which amounts to the same thing, if an average of hundred output units come forth during the period, the error which is introduced by the definition of output as units completed can safely be neglected: it is not likely to obscure variations deriving from other sources. If the proportion is about fifty per cent, the researcher must revise his definition of output. The most obvious revision is to extend the time period sufficiently to make the error negligible. Output could also be defined as the inverse of the time taken to complete one unit of product. When it is not necessary to measure various operators' outputs for exactly the same time periods, this method may prove very useful. When it is necessary to do so, this method cannot be used: prorating of output for each time period requires the assumption that the operators' efforts remain constant from one time period to the next, while the research project would require measurements of variation in efforts from period to period and would thus not permit such an assumption.

Direct method of measurement

After defining the term "physical output at the job level" the researcher can decide on the method of measurement. Physical output can be measured either directly or indirectly,

[8]Rothe, Harold F., "Output Rates Among Butter Wrappers II," Journal of Applied Psychology, Vol. 30, No. 4, August, 1946, pp. 320-327.

[9]These definitions are very similar in one sense: by adding the output rates of thirty-two subsequent fifteen-minute periods one obtains an eight-hour output rate, by adding five subsequent eight-hour output rates one obtains a forty-hour-week output rate, and so forth. But very different measuring techniques would ordinarily be used for different output periods. In most cases it is superfluous and unnecessarily expensive to measure fifteen-minute variations if one is interested only in daily variations. Also, if one is interested in daily variations a measure of weekly output alone is useless since averaging through division of the weekly output by five, by its very definition, eliminates the variations. Therefore, definitions using different length of the time period could be treated as somewhat different definitions.

and several different techniques can be used in both cases. The direct method consists in counting, weighing, or measuring the physical goods turned out. The indirect method consists in counting, weighing, or measuring some one other variable that predicts with sufficient accuracy the variable one would like to measure, namely physical output. In both cases there are problems of validity and reliability of the special techniques to be used. Validity is the extent to which a measure does measure what it is supposed to measure. Reliability is the consistency with which a measure does measure whatever it measures. A measure may have very high reliability despite very low validity. For example, two successive readings of the weight of a group of individuals may show very high consistency; yet they do not measure their intelligence very well. But a measure cannot have high validity without having high reliability; if two successive readings of the weights of a group of individuals are very inconsistent, at least one of the two sets of readings cannot measure accurately the weights of these persons.[10]

Since physical output is the variable that one wants to measure, it would seem that the direct method of measurement must be valid by and in itself, that is, it would have "face validity." Such is, however, not necessarily the case. The following examples may make this clear. In a study of the physical output of blankets,[11] existing records of output had to be utilized. The routine by which the records were established was as follows. Power sewers recorded daily output of each size and type of blanket on the back of their daily time cards. (They obtained their data from the tags attached to each batch of cut blanket cloth). The time cards were collected and reviewed by the foreman who decided whether the data seemed reasonable and who, after occasional adjustment together with the employee concerned, turned in the cards to the payroll department. Since the foreman did not apply any accurate over-all check of reported output, it might have been possible for one or more sewers to report consistently a higher output then actual; whether this really happened, the writer does not know. Under this reporting system the operators could also report less than their actual output one day and report the difference some other day. According to the foreman this was actually done by some operators.[12] While measurements of monthly output were hardly affected to any significant extent by this practice, the measurements of daily output were affected: in several cases they showed little variation although actual output no doubt varied more. Of course, this practice will impair desired research on daily variations, but the actual amount of distortion produced thereby remains to be measured.

In Rothe's study of butter-wrappers[13] output measurements were obtained by the investigator and his assistants in what appears to be a more accurate way. Mechanical counters were placed along the wrapping table, one counter for each wrapper. Whenever an output unit was completed, the operator depressed a lever. Every fifteen minutes the investigator and his assistants read the counter readings, and recorded the output. Considering that at least one investigator was always present and that the counter readings had nothing to do with the operators' earnings (hourly rates were in effect), one may expect very high validity from this procedure. Furthermore, the butter wrappers were accustomed to the "counter measurement" as a regular management procedure introduced from time to time in the past.

---

[10] More exhaustive discussions of concepts and measurements of validity and reliability can be found, for example, in Thorndike, Robert L., Personnel Selection, New York, John Wiley and Sons, Inc., 1949, pp. 68-185.

[11] This unpublished study was conducted by the writer as a research assistant in the Industrial Relations Center at the University of Minnesota.

[12] For a description of similar practices in other plants, see Gardner, Burleigh B., Human Relations in Industry, Chicago: Richard D. Irwin, Inc., 1945, pp. 150-162.

[13] Rothe, Harold F., "Output Rates Among Butter Wrappers I", Journal of Applied Psychology, Vol. 30, No. 3, June, 1946, pp. 199-211.

Since direct measurements may have less than full validity, one may ask whether and how the validity of such measurements can be measured. There is probably no single answer to this question. When the study concerns past events no new records can obviously be made; those that may exist were probably designed to serve payroll and cost-accounting needs, and the checks available often cannot be expected to be accurate enough for research purposes. When the investigator makes his own measurements, albeit through employee reports as in Roth's study, spot checks of the validity of fifteen-minute output measurements may be made by inspection, and these additional short-period measurements may be compared to the regular readings for the same periods (counter readings in our example) and validity coefficients may be obtained by means of correlation. For a rough over-all check one may accumulate the regular short-period measurements and check their daily totals against daily totals measured in other parts of the production process.

Some unreliability of direct measurements may be expected.[14] In the study of existing data on output in power sewing, unreliability may have resulted from errors in recording the batch contents on the tags and from operators' errors in reading the tags, in adding the figures and in recording the sums on the time cards. In the butter-wrapping study unreliability may have resulted from the operators depressing the counter levers too few times or too many times or from faulty reading and recording by the investigator. Among the common methods of measuring the reliability of psychological tests, namely the test-retest method, the equivalent-forms method, and the split-half method[15] only the equivalent-forms method appears theoretically applicable to the measurement of this type of reliability. If used in practice this method would require that a second investigator sample a certain number of the time periods that are also observed by the regular investigator, and measure the output in each of these periods independently of the latter investigator. This method can naturally not be applied to data collected in a past period; hence the reliability of such data will remain unmeasured. How disadvantageous this is depends not only on the magnitude of actual reliability but also on the needs of the study.

What has been said above regarding validity and reliability of direct measurements refers not only to errors in measuring quantity of output, with time limits assumed to be observed strictly, but also to observance of these time limits. It applied as well to product identification data in multiproduct operations.

Indirect method of measurement

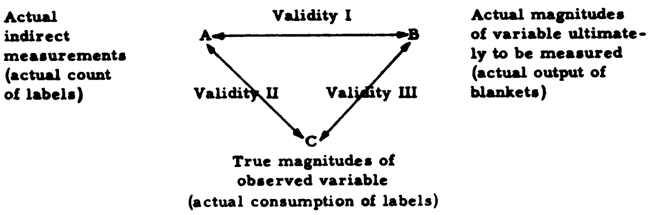
Indirect measurements are taken when direct measuring would be highly impracticable or too expensive in relation to funds and resources available or to desired accuracy of measurement. As in the case of the method of direct measurement accuracy is measured in terms of validity and reliability. In the case of the indirect method of measurement there are, however, three concepts of validity instead of one. The first concept of validity (validity I) is analogous to the validity concept of direct measurement; it refers to the extent to which the measurements measure what they are intended to measure, that is physical output as defined. The second concept of validity (validity II) refers to the extent to which the measurements measure the variable under direct observation. The third concept of validity (validity III) refers

[14] It should be noted that this concept of the reliability of a direct measurement differs from a more usual concept of reliability of output data. While both concepts arise from recognition of sampling errors, the latter concept refers to the consistency with which a time sample measures the output forthcoming during the time period that constitutes the time population. The former, on the other hand, refers to the consistency of sample measurements for one and the same period of time, drawn from an hypothetical infinite population of measurements for that time period. A similar distinction can be made with regard to the validity concept. Problems of time sampling are discussed below.

[15] See, for example, Guilford, J. P., Psychometric Methods, New York: McGraw-Hill Book Co., 1936, pp. 411-413.

to the extent to which the true measure (error-less measure) of the variable actually measured does measure whatever one wants to measure ultimately, that is physical output. The following graph and example may make these distinctions clear.

Graph 1
Three concepts of validity



In a second phase of the above-mentioned power-sewing study an attempt was made to measure physical output for fifteen-minute periods. Fourteen operators were scheduled to work as power sewers, and only one recorder could be made available. It would have been too time-consuming to have the recorder count the completed blankets directly. The following simplified method of measurement was finally devised.

Since a label indicating the company's name was to be sewn on to each blanket in the last phase of the sewing or stitching operation and since some operators had long been used to counting up labels in advance, in order to "keep track of their output", all operators were instructed to count up piles of ten labels each and to use up the labels in one pile at a time. The recorder could then count the number of remaining labels very easily, and the investigator simply computed output from the recorded pile contents.

This is a clear case of indirect measurement. Since there was supposed to be one-to-one correspondence between the number of blankets completed and the number of labels used up, the latter number would provide a convenient measure of physical output. But was this supposition factually correct? In other words, was the "true" measure of labels used up a perfectly valid measure of output, that is, was validity III perfect, with a correlation of 1.00 between B and C (see graph 1)? The answer would be no, if some types of blankets regularly did not carry labels in their finished state or if some types of blankets regularly carried more than one label each or if some labels were defective and had to be discarded. Such, however, was not the case; the true measure of label consumption was therefore a valid measure of output.

The actual measure of label consumption could, however, still differ from the true measure, both systematically and erratically, that is validity II could be less than perfect. Conceivably the operators could have counted up eight or twelve labels at a time without the recorder detecting this. Or the recorder could have mis-counted the number of labels remaining in the piles. If this were found, systematic deviations could occur, and the actual measure would not be perfectly valid with regard to the true measure of label consumption.

The three validity concepts can now be explained with reference to actuality. Validity I refers to the accuracy one would achieve in measuring actual output of blankets by counting the number of labels used up <u>if this counting could be done with no error</u>. Validity II refers to the accuracy actually achieved in counting labels used up. Validity III refers to the accuracy achieved in measuring actual output by actual counts of labels used up.

Now, what is the usefulness of the concepts termed validity II and validity III? Why is the first concept, validity I, not sufficient? It is true that validity I is all that is needed if one wants merely to know how well the obtained indirect measurements do predict actual output. However, if one finds that measured validity I is not satisfactory, one will want to know where the sources of error lie. The concept, validity II, then focuses attention on the errors made in counting the number of labels remaining in the piles, in computing consumption of labels, and in counting up ten labels for each pile. The concept, validity III, focuses attention on lack of one-to-one correspondence between the number of blankets completed and the number of labels actually used up. These latter two concepts thus lead to methods of measuring separately errors deriving from two distinct sources and of improving measuring techniques.

The techniques used in measurements of validity in the indirect method are no different in principle from techniques used in measuring validity in the direct method. In the second phase of the power-sewing study, a reasonably close approximation to validity III could be obtained by having an extra observer sample finished blankets to find whether the blankets carried one and only one label each. Validity II could have been measured by having two independent observers, one who counted the labels on the power sewer's desk, and one who counted actual consumption of labels directly, both for the same sample of time periods, and by comparing the two sets of records. Validity I could also have been measured by having two independent observers. One of these observers would have counted the number of labels in the piles at the beginning of each of a sample of time periods, and the other would have counted the number of blankets completed during each of the same sample of time periods. Validity I would then be measured as the degree of correspondence between these two sets of records.

Indirect measurements may, of course, also be unreliable. In the above study of fifteen-minute rates of blanket output unreliability of measurement might have resulted from labels being lost, from mistakes in counting on the part of the power-sewers, and from mistakes in counting and recording on the part of the recorder. The method of measuring reliability of indirect measurements is analogous to the method of measuring reliability of direct measurements described above.

As in the case of direct measurements validity and reliability are also affected by errors in the observance of the limits of the time periods, and in the recording of product identification data in multi-product operations.

Sampling of time periods

The preceding discussion of direct and indirect methods of measuring physical output is concerned with errors in the individual measurements themselves. No attention was paid to the question whether and to what extent measurements of physical output for particular periods can predict physical output in other periods. The errors that were considered are sampling errors to be sure. But the methods of sampling were not methods of sampling time periods but methods of sampling measurements of physical output for given time periods and given operators no matter how these time periods or operators were chosen. Attention can now be given to the question how these time periods and operators should be chosen in order to permit the investigator to draw inferences from outputs for time periods and operators that were observed, to outputs for time periods and operators that were not observed. The present problem is, in other words, one of sampling time periods and operators. The problem of sampling time periods must be chosen with relation to the purpose at hand. The problems connected with such methods appear to be most easily discussed with reference to such problems.

Let the first problem be the measurement "long-range" output of an individual employee. Such measurement may be made for several reasons, one of which is the validation of selection tests.[16] Since it is usually impossible, for practical reasons, to ascertain the output of an employee for his total time of employment with the firm, one must decide on the length of the time period that will represent this long-range performance. The question then arises how long this time period should be. The answer clearly depends on the amount and kind of variation that physical output shows. If the daily average of hourly output is always the same, it suffices to measure one day's average hourly output. Such a case would, however, be very exceptional. One may expect some variability from day to day, some variability from week to week, some from month to month, and also some variability from year to year. The time period chosen as representative of long-range output becomes a sample from a population of possible time periods. Is there a method of estimating the error involved in using this period as "representative" period?

[16]A selection test is, of course, not validated on one employee's output; a whole group of employees is needed. Validation on the output rates of several employees may be done only insofar as the output rates of individuals can be compared. Comparability of output rates is discussed briefly below.

If variations over the course of time were only of a random nature, fiducial intervals for "true" long-range output could be established on the basis of the variability within the "representative" period. This case being unlikely, one may study the variations over long periods of time for persons who have been employed by the firm a long time, and some conclusion might be drawn as to the error limits for any chosen length of the time period. If this is not feasible, one must rely on pure judgment as to the representativeness of the period.[17]

Once the specific time period is determined, one must decide whether to measure the output for the complete period or to secure a sample estimate of the output period. Frequently measurement of output for the complete period is preferable to measurement of output for sample parts of the period, since the whole population of parts is then measured and since there are therefore no errors of time sampling. It may, however, be too expensive in relation to the degree of accuracy that is needed; then the sampling method is preferable. Complete enumeration may, on the other hand, yield less accuracy than the sampling method; when the population is very large, one may be forced to resort to cheaper equipment or labor, and monotony in taking large volumes of observations may lead to inattention and resulting errors of measurement.

When the output of the whole period is to be measured and when only one product is turned out in the observed operations, no technical problems appear to be present beyond those mentioned in the above discussions of definition and of direct and indirect measurement methods. In the case of multi-product operations one serious difficulty is likely to appear when one intends to use past records. The time spent by an operator on each of the various types of products is not likely to be available in past records, especially if these are designed to serve payroll purposes only. When a common denominator for the various types of product can be used, this fact presents no difficulty: operations in which this applies can, in fact, be considered as single-product operations. When there is no acceptable common denominator, the total population of time periods must be viewed as an aggregate of several populations, one for each type of product, and one for each combination of types. All time periods during which product A alone was turned out, make up one of these populations; all time periods during which product B alone was turned out, make up a second population; all time periods during which both product A and product B were turned out, make up a third population; and so forth. The output of each product can then be ascertained separately while neglecting the time periods during which other products were turned out. But this method of measuring the output flow can be applied only when the type of output during each time period is known. When existing data are used, this usually means that the sample time period has the length of at least one day. A considerable loss of data then follows and when a large number of types of products are turned out, the study may have to be given up due to lack of a sufficient number of useable time periods. This is even more true when time data are available only on a weekly basis.

The difficulty of obtaining sufficient data therefore makes it advisable for the researcher whenever possible to collect his own data or to have the firm rearrange its output and time statistics so that they may be used in the study.

When the researcher collects his own primary output data, he may often prefer to use some sampling technique instead of measuring the whole "representative" population. Shortage of personnel and funds may produce this decision; it may also result from a judgment that the sampling errors that will occur, will remain within acceptable limits. Techniques for sampling within the "representative" period may also be found useful when

[17]Among the reasons why this might not be feasible are lack of output records covering a sufficiently long period, changes in the products turned out in the operation, and changes in the operation itself.

the researcher is dealing with existing data.[18]

Sampling techniques for measurement of output would be simpler in single-product operations than in multiple-product operations. Techniques for the former procedure can easily be obtained by simplification of techniques used in the latter procedure; consequently, only techniques for sampling output measurements in multi-product operations will be discussed.

It is almost inconceivable that the proportions of the work day that an operator in multi-product operations spends on each of the products, vary from day to day in a random fashion. On the contrary some products have "long runs", others come in small lots.[19] In order to obtain representativeness under such conditions one would have to draw a random sample of days from the total period that is taken as representative of long-range output. The rather common technique of drawing every nth unit from a given population may lead to sampling bias; this is particularly clear when every fifth day is selected from a five-day work week, since there is reason to expect systematic differences in output for different week days. One can either select an n which differs from the number of work days per week, or one may use a table of random numbers for selection of days.[20]

How many days out of a "representative" period one should sample to obtain a sufficiently accurate estimate of the output in the "representative" period is difficult to say without experimentation. The larger the number of types of products, the smaller is the chance of selecting a day when a given type of product is being turned out, and the larger number of days must one select in order to reach a sufficiently accurate estimate of the output of that product. Moreover, the larger the variations in output from day to day, the more days must be selected to yield sufficient accuracy. Sampling theory can no doubt be applied with profit to the determination of an adequate sample size.

The preceding discussion of problems of sampling time periods was concerned with measurement of "long-run" physical output of an individual employee. When one wants to measure "long-run" physical output of several employees, as is always the case in validation of selection tests, the difficulties are increased. Since, in the case of test validation, one wants to study the relationship between obtained test scores and actual output, measurements of output of members of the group must be comparable. Such comparability may be very difficult to achieve in multi-product operations. It was noticed in the power-sewing study that some operators were regularly assigned to stitch one or a few types of blankets while other operators handled other types of blankets; some types of blankets were handled by all operators, but this occurred infrequently.[21] If the "standard blanket" construct could not be applied in this situation, comparison of the operators as to their output rates would be difficult. A large sample of time periods would be needed to produce the required number of observations of output of each selected type of blanket; a sufficiently large sample could perhaps not be obtained. But even if this first requirement is met, that is the operators' output consists of a common type of

[18] Most firms with single-product operations are likely to have both time and output records at least on a weekly basis. Such data suffice for measurement of long-range output. When the concern is with variations within a week or within a day, such data are less likely to be available.

[19] It is perhaps equally improbable that one will find multi-product operations in which all operators performing the same kinds of operations show the same composition of types of products. The additional complications that this introduces are discussed in connection with problems of comparing the physical output of different operators.

[20] Such a table of random numbers is available in Snedecor, George W., Statistical Methods, Ames: Iowa State College Press, 1946, pp. 10-14. Applications of sampling by using tables of random numbers to time-study problems are reported in Barnes, op. cit., pp. 384-387.

[21] The foreman attributed this phenomenon to ability differences between operators and to personal preferences for some types of blankets over and against other types that were actively disliked.

blanket, the output data may not be comparable. Since there appear to be systematic variations in output rates and perhaps something approaching periodic variations, obtained data must cover peaks and troughs in output as well as intermediate stages. It therefore appears necessary to require, in addition, that measurements be spread rather uniformly over the total "representative" period. If these two requirements cannot be met at least with a fair degree of approximation, one would have to give up the attempts to compare output rates of various operators; when the ultimate aim of the study is to validate selection devices, this means that such devices cannot be validated against output and that substitute validity criteria like output ratings must be utilized.

A third illustration of problems of time sampling may be given by reference to a hypothetical study that attempts to measure the effect of introduction of a piece rate system of pay upon the output of a group of factory workers.[22] It is possible that the effects of the piece rate system, if any, show up in output only long after the system is introduced. If one wants to find not only whether a change in output actually takes place, but also when it takes place for each individual operator and the extent to which each individual operator's output changes, the measurement problem becomes quite complex. First of all, it is necessary to include a total time period long enough to take care of "delayed reactions." Second, the output data must have inter-individual comparability. Third, the total time period must be covered rather thoroughly; otherwise one may not be able to determine the times when changes do take place. In multi-product operations these conditions often cannot be met; this applies particularly to situations where existing data referring to a past period must be used. A before-and-after study of output rates probably requires that the investigator establish his own method of measuring output, before the change in the system of pay is made.

Sampling of employees

A few words may be appropriate about sampling of employees. In most cases so few operators perform essentially the same operations that sampling is completely unnecessary; the problem usually consists in obtaining a sufficient number of operators to study. Although there may be numerous operators performing essentially the same operations, these operators may, furthermore, turn out such a variety of sizes and styles of products (such as blankets or shirts) that it is inadvisable to exclude any of the operators. If sampling of operators is considered necessary despite these considerations, two methods of random sampling appear to apply. In one method, a list of all operators in the population is prepared and persons are selected at random, either by taking every nth person or by using a table of random numbers. In the other method the shop or department may be divided into areas; some of these areas are selected at random, and the operators having their work places in any of these selected areas are included in the sample. Both methods take account of the possible non-random placement of good and poor operators within the shop or department.

Suggested Contact Procedure for Outside Research Workers

The following suggestions concern the appropriate contact procedure for outside research agencies and independent research workers. They are based on the writer's limited experience in establishing research contacts and should not be taken as definitive.

A clear and concise plan of the study should be presented to the officials of the firm in their own language. The plan should emphasize both the research aims of the study and the values that the firm may derive directly from the study; it may not be necessary to present all the details of the study in this plan or in interviews with the officials.

[22] The design of this experiment will not be discussed here. If association is found between introduction of a piece rate system on the one hand and changes in output on the other hand, there may or may not be a causal connection. Control of additional variables is as usual a most difficult problem.

Top officials of the firm should be contacted first: in addition to having more authority than lower officials, they are more likely to take a positive attitude toward a well-designed research project.

When contacts with lower officials are subsequently made, less information should be volunteered about the purposes of the study and emphasis should be placed on gathering information relevant to the design of the study.

If the investigator plans to collect output data by direct observation, rapport must be established with the operators. If the operators are represented by a union, the shop steward or a union official should be invited to join the supervisor or the official of the firm in announcing the study to the operators.

After the data have been collected, the investigator should prepare a preliminary report to the firm on data or findings that may have immediate value to that firm, yet without revealing the operators' identity. When possible, preparation of this report should precede detailed analysis of obtained data.

It is almost necessary to construct in advance a check list that covers all required information on availability of records, methods of data collection utilized by the firm, flow of material, layout of the work place, job descriptions, etc. A well-designed check list built on research hypotheses and planned design is the best guarantee that required information will be gathered. The writer hopes that the preceding discussion of measurement of physical output will prove helpful in constructing such check lists as well as in designing the output study.

Publications of the Industrial Relations Center

RESEARCH AND TECHNICAL SERIES

Accounting Methods for Local Unions. *Research and Technical Report 1.* March, 1950. **$1.00.**
How to Make a Wage Survey. *Research and Technical Report 2.* March, 1950. $1.00.
Employee Handbook Printing Practices. *Research and Technical Report 3.* April, 1950. **$1.00.**
Conducting Wage Surveys. *Research and Technical Report 4.* February, 1950. $1.00.
Minnesota Manpower Managers. *Research and Technical Report 5.* February, 1950. $ .50.
Use of Factorial Design in Industrial Relations Research. *Research and Technical Report 6.* November, 1950. $1.50.
Employee Welfare and Benefit Programs: Proceedings of a Conference Held March 23-24, 1950. *Research and Technical Report 7.* November, 1950. $1.00.

How to Build a Merchandise Knowledge Test. *Research and Technical Report 8.* July, 1951. $1.00.
Job Evaluation Practices. *Research and Technical Report 9.* July, 1951. $1.00.
Measurement of Physical Output at the Job Level. *Research and Technical Report 10.* July, 1951. $1.00.

Orders for Research and Technical Reports should be addressed to:

WM. C. BROWN CO.
915 Main Street Dubuque, Iowa

BULLETIN SERIES

Yoder, Dale, D. G. Paterson, et al. *Local Labor Market Research.* 1948. $3.50.
Training and Research in Industrial Relations: Proceedings of a Conference Held May 25 and 26, 1945. *Bulletin 1,* October, 1945. Single copy, $1.50.
Labor-Management Cooperation: A Case Study in the Minneapolis Laundry Industry. *Bulletin 2,* July, 1946. (Out of print.)
Jobs in Industrial Relations: A Description of Some Principal Industrial Relations Jobs in Labor and Business Organizations (with charts). *Bulletin 3,* January, 1947. Single copy, $ .75.
Training and Research in Industrial Relations (II): Proceedings of a Conference Held May 9 and 10, 1946. *Bulletin 4,* April 1947. Single copy, $1.50.
The Industrial Relations Five-Foot Shelf. *Bulletin 5,* September, 1947. Single copy, $ .75.
Industrial Relations Glossary. *Bulletin 6,* April, 1948. Single copy. $ .75.
Training and Research in Industrial Relations (III): Proceedings of a Conference Held May 22 and 23, 1947. *Bulletin 7,* April, 1948. Single Copy, $1.50.
Jobs for All: A Primer of Theory. *Bulletin 8,* April, 1949. Single copy, $1.00.
Ten Years of the Minnesota Labor Relations Act. *Bulletin 9,* April, 1949. Single copy, $1.00.
Minnesota Manpower Mobilities. *Bulletin 10,* October, 1950 . Single copy, $1.00.
Triple Audit of Industrial Relations. *Bulletin* 11,August, 1951. Single copy, $1.00.
The Industrial Relations Center of the University of Minnesota: The Program of the Industrial Relations Center. September, 1950. No charge.
Nos. 1, 4, 7: 2 to 24 copies $1.00 each, plus postage; 25 copies and up, 75 cents each, plus postage.
Nos. 3, 5, 6: 2 to 24 copies 50 cents each, plus postage; 25 copies and up, 35 cents each, plus postage.
Nos. 8, 9, 10, 11: 2 to 24 copies 75 cents each plus postage; 25 copies and up, 50 cents each, plus postage.

Orders for above publications should be directed to:

UNIVERSITY OF MINNESOTA PRESS
10 Nicholson Hall Minneapolis, 14

MIMEOGRAPHED RELEASE SERIES

Industrial Relations Positions and Personnel: A Survey of Characteristics of Personnel and Industrial Relations Officials and Union Officials. *Mimeographed Release 1.* July, 1950. (Out of print.)
Job Evaluation Practices: Proceedings of a Conference Held December 1-2, 1949. *Mimeographed Release 2.* July, 1950. (Out of print. Replaced by *Research and Technical Report 9.*)
Community of Interests Between Unions, Employers, and Investors: Proceedings of a Conference Held December 10-11, 1949. *Mimeographed Release 3.* September, 1950.
Unions in the Community: Proceedings of a Conference Held February 15-16, 1950. *Mimeographed Release 4.* September, 1950.

Mimeographed Releases may be obtained free of charge from the Industrial Relations Center as long as the supply lasts.

REPRINT SERIES

Salaries and Manpower Managers. *Reprint Series: No. 1.* (Out of print.)
Refusal Rates and Interviewer Quality. *Reprint Series: No. 2.*
Minnesota Labor Relations Acts—An Opinion Survey. *Reprint Series: No. 3.*
Validity of Work Histories Obtained by Interview. *Reprint Series: No. 4.*
Manpower Managers—Their Habits, Haunts, and Customs. *Reprint Series: No. 5.*
Personnel Salaries and Ratios in 1950. *Reprint Series: No. 6.* (Out of print.)
Readability and Human Interest of Management and Union Publications. *Reprint Series: No. 7.*

Reprints will be furnished free of charge from the Industrial Relations Center as long as the supply lasts.