

Spring 1937

Critique of the Bellman Police Service Rating Scale

Spencer D. Parratt

Follow this and additional works at: <https://scholarlycommons.law.northwestern.edu/jclc>

 Part of the [Criminal Law Commons](#), [Criminology Commons](#), and the [Criminology and Criminal Justice Commons](#)

Recommended Citation

Spencer D. Parratt, Critique of the Bellman Police Service Rating Scale, 27 *Am. Inst. Crim. L. & Criminology* 895 (1936-1937)

This *Criminology* is brought to you for free and open access by Northwestern University School of Law Scholarly Commons. It has been accepted for inclusion in *Journal of Criminal Law and Criminology* by an authorized editor of Northwestern University School of Law Scholarly Commons.

A CRITIQUE OF THE BELLMAN POLICE SERVICE RATING SCALE*

SPENCER D. PARRATT†

In his "Police Service Rating Scale," Mr. Arthur Bellman points a direction toward the increase in objectivity in measurement of police functioning. He has presented his instrument in a commendable spirit, recognizing that it is "by no means a finished product," and further that "it is only in the beginning and experimental stages, and needs to be subjected to practical use in order that further improvements may be made and its ultimate value determined." There are, however, some matters involving theoretical and conceptual considerations which appear worth presenting as commentary to Mr. Bellman's study with a purpose of increasing the final success of scaling in the general field of police administration, but which emphasize the experimental nature of the Bellman Scale as it is offered at the present time.

That Mr. Bellman is seeking to increase the realism of evaluating police administration indicates a faith in newer techniques of approach to a subject which has resisted traditional academic penetration. With any attempt to further this end the present writer is in substantial agreement, and whatever commentary follows is directed toward the attainment of the success of methods relying upon explicit data and operative through devices permitting others to follow and add further increment. Unquestionably the culmination in the creation of a successful police rating instrument will involve the interaction of many minds. What are believed to be correctable defects in the Bellman instrument are pointed out, with constructive suggestions for overcoming such shortcomings, insofar as this is possible. No conclusiveness or finality of determination is claimed and the experimental nature of the subject matter is expressly recognized.

Mr. Bellman well states the ultimate objective of a successful scaling device for measuring police administration:

"A scale implies measurement, and measurement implies certain

* Bellmann, Arthur, "A Police Service Rating Scale," 26 J. Crim. Law and Criminology, 74-114 (1935).

† Associate Professor of Political Science, Syracuse University.

established standards. No one questions present methods of measuring distance, weights and temperatures. . . . This study reports an attempt to devise standard units of measurement for police departments whereby the quality of the department can be quantitatively measured."

A basic question is presented: To what extent is the Bellman instrument analogous to a yardstick, a balance scale, or a thermometer as a measuring device?

The use of the term "quality" without explanation of connotations must be taken to mean that the thing sought to be measured is general goodness or worth in the performance of the police functions by a given department. There is an unfortunate indefiniteness in the variable thus stated. It would appear that a police department performs functions imposed from without insofar as objectives or ends of operation are concerned. Thus the extent to which the externally imposed objectives are accomplished might conceivably be made a condition upon which measurement could be projected. Or a person could define such a term as "efficiency" to provide a characteristic of performing given duties so as to indicate a generally understandable variable which might be considered subject to measurement. Or a more general phrase, as "effectiveness in performance of legal duties" might be made a somewhat meaningful variable. But the Bellman study gives no instructions as to just what is included, and what is excluded, in the term "quality" and thus introduces marginal ambiguity which somewhat defeats the purpose of the instrument as a measuring device.

Is quality of a police department like distance or time? Distance, time, or weight are abstractions. But each describes an abstract quality subject to apperception in terms of explicit data. Each requires a separate measuring instrument in terms of its particular continuum of explicit data. Quality as used by Mr. Bellman appears to be something quite different in that it describes a generalized summation of many separate abstractions. It is not the absence or presence of one thing, but the absence or presence of many things in different and varying proportions.

The significance of the above generalizations can be given greater realism through homely illustrations. A modern police department is at least as complicated as a single individual. Suppose that a list of items could be compiled by people having had considerable experience with individuals and an assigned weight given to each item. Then suppose that to the totality of all items the term quality is assigned. Would the absence or presence of a

percentage of the items, as measured through third party application of the instrument, formulate a standard of quality evaluation? Let us simplify the matter. Suppose that a group of eminent medical authorities should confer and agree that a given list of items constituted the qualitative elements in physical well being of individuals. If these items were then given weightings of two each and a qualified medical expert applied the "scale" to a given individual, it would be possible, in Mr. Bellman's method, to theoretically measure health of that individual. But what could be found out of practical value?

Application would inevitably result in reaching numerical answers. Thus A might be given a score of 60, B would get 80, etc. Such a procedure would prove academically interesting, but would be of slight value for purposes of diagnosis of qualitative differentials in health between A and B. A mere determination that a man is a given percent healthful, or that a police department is in the same predicament as to quality, is not of appreciable value to the administrator of health or police controls. Nor does it permit a legislator or voter, with such conclusion before him, to intelligently exercise a check upon administrators in either case.

Significant knowledge must be more specific. It must pick out where short-comings exist to have engineering or control value; not simply that there is abstract presence or absence of a general abstract moral characteristic. No prognosis is possible until "soft" spots are determinable. And this is impossible without knowledge of the relative values in each aspect of the total considered. Mr. Bellman makes no claim that his instrument will evaluate merits or different aspects of policing; it is limited to measuring totality of quality. As such, there is a limitation on the instrument which seems unnecessary if the instrument measures anything at all. One could hardly argue that an instrument can measure totality, but lack capacity to allocate sources contributing to that determination.

But the Bellman instrument is quite unable scientifically to claim that it can allocate or distribute weightings for the various elements contributing to police functioning. Another example will demonstrate some of the reasons. A modern police department is at least as complicated as vegetable soup. As Mr. Bellman approaches the subject of measurement of quality in a police department, he itemizes the constituent elements and presents a qualitatively selected list of such items, giving each a standardized or at least uniform value. Extent of quality is the number of items

favorably rated compared with the total number which have been listed. Suppose this process is applied to vegetable soup. One might list water, salt, pepper, potatoes, etc., and specify how much heat should be applied at what point and for what duration in time. But to obtain a qualitative formula for soup every element would have to be listed. By comparison, the Bellman instrument as formula for qualitatively describing a complete department obviously is deficient. It is, by analogy, a partial receipt, and as such cannot measure completeness under any circumstances.

Only when all the qualitative elements involved in vegetable soup are listed can any tabulation formula be concocted for the qualitative ideal. Manifestly the Bellman instrument is not of the nature of a formula for completely itemizing the correct elements in a department in their totality. It is a highly selective classification. But suppose that all of the qualitative elements had been listed. One might easily list all of the elements in vegetable soup meeting qualitative standards. By reading the literature and talking to an indefinite number of cooks, this would be possible. From this experience one could obtain formulae which involve quantitative elements. Each would be given its tested weighting for construction of soup. Mr. Bellman read extensively in police administration and talked to many experts. But no formula was forthcoming directly. He created one. Is it a tested formula for quality in police administration, or simply a list of qualities discovered to be involved without any determination of how much of each properly introduced into the concoction to be known as "high" quality police administration? The simple listing of qualitative items, plus arbitrary assignment of values to each, is no certain test that the product is able to form a yardstick to measure quality.

But what is quality in soup? Is it palatableness, nourishing or sustaining caloric contents, or yet other things? Or is it combination of these and means different emphasis and items to different men? If quality is not of such a definite meaning as to leave small marginal ambiguity there can be no scale to measure it. One cannot scramble qualitatively different units in a single variable. No instrument can measure distance and temperature simultaneously, unless it possesses independently determined characteristics for each.

Now let us assume that quality in soup is a something. Presume that it is palatableness. Another difficulty appears. How can palatableness as quality be measured so that the conclusion will

have meaning to others than the person performing the acts involved? One might say, in relation to soup coming up to a preconceived standard of palatableness, that it lacks one unit of salt per quart, contains one cup too much water per gallon of soup, or that it has been subjected to heat of a certain amount for a period of time, specified in minutes or seconds, in excess of the hypothetical standard. Whatever could be stated as having meaning common to others interested in soup is permissible in measuring quality in soup. But a statement that a given batch of soup is lacking six points in "quality" without further definition in explicit terms is quite meaningless to anyone seeking to understand anything about perfection in soup.

The same difficulties inherent in measuring quality in soup are present in measuring quality in a police department, but in an indefinitely more complex variety. Police administration is a composite of many continua, or variables, in behaviors, states of mind or attitudes, and external conditioning factors. Quality is an abstract moral term which might be significantly applied to the balance of constituent elements in what the evaluator considers correct proportions of each. A police administering system is more or less of many things, but quality is relationship between these many things in their operative conditioning. One cannot aspire to measure quality as moral abstraction without first providing for evaluating the constituent elements contributing to its totality. The Bellman instrument has undertaken an insurmountable task in disregarding the fundamental and seeking to measure the composite. It seeks to erect superstructure where no foundation has been built. The present state of police knowledge and technique calls for digging foundations.

Mr. Bellman's study should be regarded as hopeful, but not attaining its ultimate objectives. Casual inspection of data included in his itemization evidences multiplicity of variables which require evaluating separately before they can be comparatively evaluated as "somethings." Mr. Bellman makes no attempt to treat exhaustively included variables, but his schedule or tabulation of suggestive elements of importance will prove of value in formulation in these directions.

The line of attack lies in isolating numbers of significant variables involved in police functioning and treating them separately. If one cannot measure quality directly with present data, there is no reason for failure to proceed to lay the groundwork looking

toward the ultimate ideal. Once the significant variables have been isolated and measured, a foundation will be laid for a superstructure balancing their relative importance.

We have reduced the standard of quality in soup to formula because all of its constituent elements are explicitly measurable. The measuring devices are generally understood and have common meaning and as such make knowledge transferable. But without measuring units for the constituents for soup we could not formulate a standard of soup quality. We must know what the constituent elements in policing are, in explicit terms, before seeking to transcend into abstract realms of qualitative moral evaluation.

Some critical comments on Mr. Bellman's procedures in allocating his "landmark" or "weight values" to the elements in his instrument will point opportunities for significant improvement. Mr. Bellman advances an instrument containing twenty divisions. No attempt is made to guide the sympathetic follower in reaching the accorded values allowed each division. Except insofar as they appear resultant from totalling miscellaneous items at untested and presumptive values, these matters are left in mystery. The twenty main headings total 1370 points. The retention of a chief is weighted at eight points; the detective division at 376; traffic control at 56; vice at 86; patrol duties and beat construction rate 208, and crime prevention gets 74, etc. A fundamental question is raised: How can the various elements be given proper weighting in terms of the total instrument?¹

Equally wise and experienced men may differ as to the correct balancing of characteristics which shape quality or standard in any complex segment in human affairs. One might feel that crime prevention is worth as much as all other aspects of police administration together. Another might come closer to Mr. Bellman and believe

¹ Probably the nearest parallel to Mr. Bellman's instrument is the *Appraisal Form for City Health Work*, published in succession of editions by the American Public Health Association, 50 West 50th Street, New York City. (Last edition available, 1934.) This appraisal form, like that of Mr. Bellman's, includes a number of major headings. These headings have been weighted "solely on the judgment of the Committee and its many advisors," p. 4. There is merit in this procedure in constructing an appraisal form because it provides consensus of expert opinion. The Probst Service Ratings Plans (Service Ratings, *Technical Bulletin* No. 4, 1931) are also suggested by Mr. Bellman's instrument. Probst, like Bellman, does not evidence the process by which his weighting of items is determined, but he includes suggestive checking and testing methodologies. Probst recognizes variables and accords values, but his method is obscure.

that the great weighting of characteristics in police administration should go to enforcement and beat patrol.²

There are at least two defensible methods of securing weighted values for major headings in a complicated segment of related factors. Assuming that a qualitative enumeration of headings is determinable, it is possible to observe actual sequences of behaviors in relation to their external conditioning, and experimentally determine which sequence best attains ends externally imposed upon a police department. This is very difficult and quite outside of the scope of a single investigator, or even a staff of investigators, without discretion to control conditions involved in police functioning. The second method of ascertaining the relative importance to be accorded to particular qualitative segments in police functioning is through concurrency of expert opinion. Although this is less defensible as "scientific" in its procedures, it has the merit of attainability. The fundamental difficulty inherent in the second method is that it really averages values of a list of persons, who, no matter how experienced or trained, all operate under similar environmental backgrounds and in situations where exchange of ideas tends to produce stereotypes and prejudices common to the group. Moreover, it is difficult to ascertain experts in the sense of persons possessing extraordinary insight into institutional understandings. But either method is better than arbitrary weighting or mystery.

The same type of criticism and suggested means of improvement applies to the break-down of major headings into component parts. For example, the three "E's" of traffic control are weighted: Enforcement at 24 points, engineering at 18, and education at 14. Why is engineering worth four more points than education, but six less than enforcement? Or again, combatting gambling is rated at 26 points, with equal weight given to combatting the illegal sale of liquor and two additional points for combatting prostitution. Why? Perhaps the instrument has some undisclosed method of attaining these weightings; perhaps the weighting is fortuitous in terms of discoverable unit items entered under each.

The same problem appears in equally acute form in considering the individual itemizations. There are 685 of these. Manifestly this is a selected list and could be increased to untold thousands by

² Bellman gives 376 to the detective division, 156 to patrol duties, 52 to beat construction, or a total of 584 of 1370 units. This is over one-third of the scale values. To crime prevention, Mr. Bellman gives 74 points, although a variety of points are related to this phase of police activity but scattered in other places in the instrument.

any scholar or practitioner of police administration. Just how did the selection of the relatively small number take place? The author indicates the range of his experimental conditioning in prefacing his instrument but does not produce explanation for the selection of his items. Moreover he recognizes that "all individual items listed in the score sheets are not of equal weight." But why should they be given equal weight if this is true? There appears no evidence that errors will be averaged out in application. If they are cumulative in their effect, the toleration of resulting error is inexcusable, at least when there are easy methods of eliminating such difficulties available in the data on the subject of scaling and scientific techniques.

Why did the weighting of two points come to be assigned to each item? Some of the particular items may be worth, since no test of weight values has been made, ten or a hundred times others. The number "two" may not permit adequate qualitative evaluation for particular differentials. Really the author is involved in formulating a three point scale on each statement; from 0 to 2, and in this respect the instrument is not a completed scale, but rather an un-evaluated list of items which might be made into a scale if opportunity should be offered and other defects cured.

A student of police administration seeking to improve the Bellman scale in relation to the weighting or evaluating of relative items included as qualitatively essential to description would need only to look to established methodologies for assistance. The development of the Thurstone technique of scaling is directly in point. Although Professor Thurstone applied the device to attitudes it would appear useful and competent as a method of scaling values of practices in terms of determined variables. By analogy the same work methods would apply to weighting or evaluating within a range of variables in the same datum plane, either to factual conditioning or status arrangements, or both. The fundamental ideology involved in accomplishing such results centers around the attainment of concurrencies of agreement among persons selected as desirable or competent to meritoriously balance and evaluate in terms of institutional understandings. Thus, in evaluating or weighting items in a police service scale, it would appear eminently practicable to obtain the evaluation of a group of experts through a device permitting accurate tabulation and formulation of extents of concurrencies. Mr. Bellman has not even indicated a knowledge of the problems in-

volved in the rather extensive writing on the subject, although such materials are readily available.

The great merit of a valid scale is that it relies upon data subject to observation and verification. This is what gives it realism and takes it out of the realm of the metaphysical and as such offers a door to a more scientific approach to administration as process and procedure. But Mr. Bellman does not quite make a necessary step in formulating his items. Many of the included items are subject to observational verification. Others are not. It is possible to state all measurable items in terms of tangible indications.

A few examples, taken from the instrument, will show the nature of this difficulty. Under the first major heading, "Selection of the Chief," the first recorded item is factually answerable and can be reduced to a simple yes or no. It is: "Is the chief selected by the mayor or other executive head of the city?" Such a question is answerable by observation and checking of records. It does not permit of a three point scaled answer as to more or less of presence or absence. It is either positive or negative or a combination of elements making the question unanswerable as framed. The second scored question is qualitatively different from the first and asks an evaluation of motive or state of mind of an official or combination of officials. It is: "Is the selection free from political considerations?" This is a matter of opinion, depending upon the meaning of the words involved in the statement which include noticeable ambiguity. What is a "political" consideration? It is proper that any executive or high administrative policy formulating official be chosen for reasons which might be termed political if the ambiguously undesirable connotations are excluded. Such a question is not subject to observational verification by even the most competent observer. But it is possible to observationally verify whether or not past practices or behaviors involve sequential relationships from which a defined quality might be ascertained to a degree of accuracy that successive individuals might come to a high standard of concurrency of agreement with the same data. Mr. Bellman's third question is not factual in its nature, but phrased to include a range of data beyond observational check. It is: "May non-residents of the city be appointed?" Such a question involves answering a hypothetical condition outside the scope of reality and as such it is not ratable in terms of actualities. The questions, "Have non-residents of the city been appointed?" or "Does the law permit non-residents of the city to be appointed?" are subject to observational verification from record sources.

Perhaps there is no necessity of multiplying illustrations of the confusion evidenced by the particularized itemization. Before any adequate measuring device can be formulated it is essential that data be selected from a single datum plane or from the same characteristic level of knowledge. This does not mean that only practices and procedures are subject to scaling instruments. It is quite possible to take a level of factual data describing the conditioning of the department as an operating agency. Thus it would be possible to make a comprehensive, or well selected, list of facts involving facilities, equipment, physical surroundings, levels of intelligence, of appearance, or of any of the items listed as "objective measurements" by Mr. Bellman at the end of his instrument and ascribe weighted values to them.

Similarly it is possible to list and classify opinions or attitudes involved in the policing function, held either by officials or by private citizens concerning the enforcement process, and reduce them to measurable scales in terms of known and clearly determined variables. There is sufficient experience in accomplishing such objectives to take the matter outside of the sphere of speculation, at least until new and different techniques are developed to supersede those now in use. But before any attempt to combine essentially different types of data requiring quite different treatment into a common instrument can be made, it is essential to carefully test and evaluate the efficacy of the different variables involved.

Summarizing the shortcomings of the Bellman instrument to measure the functioning of police departments, and indicating specific remedial suggestions, produces the following:

I. *Conception of Data.* The Bellman itemization is highly selective, rather than comprehensive detail. As such it requires explanation as to the procedures and justification involved in selection to warn any person seeking to apply it as to its limitations. There is confusion in including questions seeking qualitatively different data for answering. Some questions ask for explicitly denotable fact, others for verifiable or at least measurable opinion, and yet others ask imponderable queries through confusing implicit and explicit conclusions. This type of defect is easily correctable by keeping criteria of data characteristics in mind in formulating items. The use of the statement form rather than question form permits sharper differentiation of meanings and this quality appears desirable.

II. *Clarification of Variables or Continua.* There is failure to appreciate that no instrument can measure more than one continuum

at a time. "Quality" is so vague and undefined a variable as the "scale" is presented that it is questionable as to just what is undertaken to be measured and in terms of what standard. Measurement, in any complex situation, can be done only through explicitly denotable data in terms of institutional understandings as evidencing "more or less" of a single continuum. The suggested approach is analysis of complex data for clarification of included variables reducible to institutional knowledge by understandable continua. These variables may be "more or less" of attitude, importance of fact to accomplishment of end, existence of particular conditioning, either physical or mental, in terms of knowable extremes. But more or less of abstract moral quality is meaningless for any engineering use or administrative reliance.

III. *Weighting of Significant Elements.* A measuring device implies a standard of something against which a given differential in related data can be projected and similarities and differences demonstrated in terms having common meanings. The idea of more or less in terms of a standard implies known values for such standard. Since these values must have generally understood meanings, they cannot be arbitrarily determined. Thus, it is indefensible to divide a measuring device into arbitrary or otherwise unexplained divisions and claim that common meanings inhere in such qualitative units in quantitative terms. This criticism applies to the weighting of major headings, sub-headings, and particular questions in the Bellman instrument. The defect can be corrected by following established methods of weighting values, either of measuring agreement among expert's opinion or by obtaining some type of group approval of assigned values.

IV. *Testing Instrument for Reliability.* An untested instrument is not ready for use outside of a laboratory. The Bellman instrument is no more than a qualitative list of significant items assigned arbitrary values and ordered into larger divisions having whatever weights happen to cumulate at each heading. There is no excuse for failure to test ambiguity in meanings of divisions, sub-divisions and particular statements in terms of institutional or generally accepted meanings. Until these things are done the instrument is a crude appraisal form or schedule and not a measuring device of any repute. The literature overlooked in preparation of the Bellman product adequately explains the procedures and methods to rectify these shortcomings.³

³ "Measuring the Severity of the Third Degree," 24 J. Crim. Law and Criminology 485-503 (1933) and bibliography footnoted.