

The Ecology of Assessment: Evaluation in Educational Settings

Robert J. Wilson

Ruth Rees

queen's university

A theoretical model of evaluation for complex educational settings would be welcome. We propose a model of evaluation with three components (measurement, judgement, and decision making) with related criteria (reliability, validity, and utility), and another for educational settings based on the work of Parsons (1986), Beer (1979), and Rees (1983). The latter model depicts the provincial educational setting as hierarchical and interdependent with lower levels nested in the operational subsystem of levels immediately above them. The development of a provincial item bank illustrates an application of the models.

On a grand besoin d'un modèle théorique qui servirait à l'évaluation de cadres éducationnels complexes. Nous proposons un premier modèle qui comprend trois composantes (mesure, jugement et prise de décision) et des critères complémentaires (fiabilité, validité et utilité) ainsi qu'un second modèle fondé sur les recherches de Parsons (1986), Beer (1979) et Rees (1983). Ce dernier modèle décrit le cadre éducationnel provincial comme hiérarchique et interdépendant, les niveaux inférieurs étant situés dans le sous-système opérationnel des niveaux immédiatement supérieurs. Le développement d'une banque d'items provinciale illustre l'utilité de ces deux modèles.

In evaluating aspects of public education, expensively developed instruments intended for one purpose, such as student achievement, are frequently turned to others, such as candidate selection, for which they have little if any demonstrated utility. The use and misuse of examination results from the final years of secondary schooling are obvious examples of this tendency. One reason for this pattern is that there are different levels in complex educational settings, and each level uses the information it has or can acquire whether or not the data were produced for its particular purposes. Flawed information is better than no information at all. Because educational levels are not discrete (or, often, discreet), tension and friction result.

It would be helpful to construct an evaluation model that could function at many levels of the educational hierarchy. Such a model would serve as a tool to describe evaluation policies and procedures used in educational structures. It could be used to locate potential gaps and inconsistencies in

present practice and to promote precise and credible assessment through greater conceptual clarity. Finally, such a model might spur discussions about its adequacy, thereby increasing our understanding of evaluation *in situ*.

A model of evaluation that recognizes the impact of various settings on evaluative practice will, then, include not only a working definition of evaluation but also an elaboration of the concepts of policy making and educational systems.

A MODEL OF EVALUATION

Most discussions of educational evaluation refer to particular locales (for example, the classroom or the province) or particular purposes (for example, achievement or selection), and models of evaluation are similarly limited.

The widest view of general purpose in evaluation is given in the program evaluation literature. For example, Scriven (1967) proposed a basic distinction in terms of goals, that Stufflebeam et al. subsequently challenged (1971). Scriven argued that the aim of any evaluation activity must be "the estimation of merit, worth, value." The alternative approach advocated by Stufflebeam et al. asserted that the fundamental aim of evaluation was to aid in decision making.

This apparently irreconcilable conflict of views may be resolved if both judgements and the decisions which emerge from them are separated conceptually. As both judgements and decisions are assumed to flow from evidence collected during a measurement phase, a comprehensive evaluation model would have these components: measurement, judgement, and decision making. These elements do not, however, exist independent of each other.

Interaction of Measurement and Judgement

In the absence of conventionally agreed-upon educational measurements with associated units and criteria, educational evaluators must construct or adapt their own for each decision. Consequently, measurement development in practice implies prior decisions about the classes of judgements it requires. This prediction of the characteristics of the results of measures means that the judgement phase includes a comparison between expected and actual results. An example from an assessment study may make this concept clearer.

McLean (1982), using basic mathematics items generated by computer-stored algorithms, apportioned six items per objective across multiple forms and disseminated these to representative samples of grades 7–10 students in Ontario. Consider each set, with accompanying difficulty levels (*p*-values), to represent a homogeneous set of items, about as homogeneous a set as one might hope to obtain. These are relatively unambiguous, rule-derived ex-

TABLE 1
*Grade 9 (Advanced) Results on Six Items Measuring
 Multiplying Two Mixed Numbers**

<i>Item Number</i>	<i>Item</i>	<i>P-Value (approximate)</i>
1	Simplify the following expression: $1\frac{3}{4} \times 2\frac{1}{2} =$	.53
2	Simplify the following expression: $2\frac{3}{4} \times 1\frac{2}{3} =$	.67
3	Simplify the following expression: $3\frac{1}{4} \times 2\frac{2}{3} =$	.47
4	Simplify the following expression: $1\frac{1}{5} \times 2\frac{3}{4} =$	.25
5	Simplify the following expression: $1\frac{4}{5} \times 3\frac{1}{4} =$	.63
6	Simplify the following expression: $3\frac{4}{5} \times 3\frac{2}{3} =$	.20

*Adapted from McLean, 1982.

amples. Table 1 displays the items and findings for the grade 9 (Advanced) sample for one such set.

One of the purposes of this field trial was to provide statistics that would help teachers judge their own students' performance more effectively. The selection of one particular item from this subset for a typical wide-ranging classroom test will be an important decision. If Item 6 is chosen, the expected *p*-value from a class typical of Ontario grade 9(A) students would be .20; but if Item 3 is chosen, that value is more than doubled and if Item 2 is chosen, more than tripled. Nor is this level of discrepancy unusual in the McLean data. Wherever ceiling or floor effects are absent, variability is common within grades and across objectives written at various degrees of specificity. In a fundamental sense, these items represent the pinnacle of possibility for homogeneity of difficulty: the scoring is relatively objective and the interaction of learner with content to produce a personal view of knowledge is minimal. In disciplines less objective than this level of mathematics, the variability of seemingly identical items would be even more extensive. By extension, a teacher could generate a series of tests in a given content domain that would produce summary statistics as wide as he or she desired. There is some evidence to suggest that that is exactly what teachers do.

In a study of one secondary school's evaluation procedures, Wilson (1989) found this pattern to be common in mark production. Typically, teachers here conducted eleven "evaluations" per student per semestered course. As part of a study of evaluation practices, the teachers reported on the purposes for each one, and virtually all the instruments had, as one of their aims, the

generating of marks for reporting purposes. Most of the instruments, especially those lightly weighted, produced the negatively skewed distributions expected with content-valid, objectives-based measures. A smaller set of instruments, often labelled “tests” or “examinations,” produced more symmetrical distributions, and this set was more heavily weighted in the final distributions.

The utility of this approach reflects an exquisitely precise understanding by teachers: if they were to meet the somewhat contradictory goals of providing feedback to students on learned materials as well as generating marks for reporting purposes on a common school-wide basis (for example, approximately as many F's as A's; medians between 65 and 72), some orchestration of the results would be required. The more frequent, less heavily weighted evaluations, then, served the instructional purposes of feedback and motivation in the classroom while the less frequent, more heavily weighted tests and examinations, with questions whose *p*-values ranged more widely, brought the final distributions into line with administrative expectations.

This example of school-based interaction of judgements and measures reflects the more general situation where expectations are established by criteria and levels outside the specific evaluation process itself. Standardized test norms, as an example of commonly available information, do not set the expected scores so much as provide a backdrop to them. If a school district in the interior of British Columbia, for example, chooses to use the *Canadian Tests of Basic Skills* to help it make decisions about programs, the Canadian norms may be helpful in judging performance, but they will not be definitive. “How well *should* our children do compared to the bulk of Canadian children?” is a problem of judgement that remains even after the norm tables are consulted. Nevertheless, some expectations are clearly built into the norms by the choice to use the tests in the first instance.

The production of local or provincial norms for tests, the comparison of these test results to others—in other words, the fleshing out of a context in which the interpretation can occur—may assist in ensuring that the gap between obtained results and expected results can be judged soundly. Still, it does not eliminate the need for judgement.

Interaction of Judgement and Decision Making

Most proposed linkages between decision making and evaluative outcomes (for example, Leithwood, Wilson, & Marshall, 1981) suggest statements of possible courses of action in advance, and connection of each set to possible combinations of outcomes. Implicit in such elaborated decision-rules are educational judgements of the worth or value of certain outcomes, judgements

which then result in decisions for substantive change, minimal change, or no change in procedures.

An example from one of the case studies reported in Leithwood, Wilson, and Marshall (1978) will illustrate this formal process. A school board was evaluating the degree to which its various techniques for implementing a set of provincial curriculum guidelines had succeeded in affecting classroom practice. The board had tried six techniques with teachers: teacher visitations, resource staff visits, provision of a teacher centre, orientation through a "live-in," staff meeting reports, and large-group sessions with speakers. Degree of participation in these activities was correlated with certain classroom indicators thought to show the orientation of the new guidelines. A threshold of perceived effect was established in advance for each strategy, and judgements about the effectiveness of the strategies were then made, alone and in combination. The application of the decision-rule followed these judgements: in this particular instance, three of the techniques (teacher visitations, resource staff visits, and provision of a teacher centre) were found useful and were used again in subsequent implementation situations while the others were abandoned.

Criteria for Effectiveness of Measurement, Judgement, and Decision Making

To be useful normatively, the model should specify criteria for each of the three components. This section will argue that the criterion for decision making is utility; for judgement, validity; and for measurement, reliability.

Reliability in measurement is equivalent to dependability, usually estimated through consistency across repetitions. One way of viewing the various reliability procedures developed for educational measures (such as internal consistency, parallel forms, inter-observer correlations) would be to relate each procedure logically to a particular array of judgements and decisions. Extending the traditional concern about instrument reliability in a single standardized setting to plausible other settings, clienteles, and their interactions, broadens the possible judgements that can flow from single instruments. These generalized reliability attributes are not yet routinely provided even though estimates for them are now theoretically and practically possible (Cronbach, Gleser, Nanda, & Rajaratnam, 1972).

Validity issues, on the other hand, have typically been bound in with purposes and judgements (Cronbach, 1971). Measures with known validity for one purpose (as in the illustration which opened this discussion) may have little validity for another. Thus, the various types of validity common to the evaluation literature (content, predictive, and criterion) can be viewed as arguments made to support rational links between measurement and certain classes of judgements. For example, the stress on predictive validity by

developers of a personality scale occurs because of the assumption that the usefulness of such a scale will frequently require judgements about the personal characteristics of job applicants, for example, in the absence of time spent in their company. On this interpretation, the broadest array of uses would be found for measures demonstrating construct validity that by definition requires determination of relationships within a network of related and unrelated concepts (Messick, 1975). Such validity would conceivably give any measure a more general applicability than that provided by any other type of validation.

Decision making moves the results from the world of deliberation to the world of action. How are we to choose which courses of action are best, especially when comparable alternatives are seldom tried? Scriven (1974) provided a taxonomy of alternatives, an array of decision choices, that reflected increasing confidence and sophistication. Recognizing the need but also the unlikelihood of "critical competitors" for many decisions, he advocated at least the conceptual comparison of educational judgements with plausible and even ideal others. Evaluating the possible choices of action according to the power of the decisions that can be made implies an ultimate criterion of utility. Decision A is better than decision B if the former is more useful.

Applications of the Evaluation Model

Given a model and its associated criteria, would test its applicability? The listing of everyday goals for evaluation—achievement, placement, diagnosis, selection, guidance, personnel, and program—perpetuates the notion of independent action. This practice has often frustrated educational implementation of evaluation policies and procedures.

Evaluation compartments and administrative levels are simply not unrelated, as the example of teachers' mark distributions demonstrates. Indeed, in provincial educational systems, evaluation policy often originates at one level with intended implementation at another. A school board developing a policy for identification of special students, for example, would expect aspects of that policy to be incorporated into the policies and procedures of schools, classrooms, and supporting professional staff.

Table 2 shows how complex and compartmentalized general evaluation policy could be in a provincial educational setting. If the purpose of evaluation is, for example, student achievement, row one of the table reveals just how many levels might be implicated in a single policy statement.

In addition, however, what is required but absent from Table 2 is a means of capturing the interdependence of hierarchical levels in a provincial educational system. Such a representation would provide a more accurate

TABLE 2
Policies, Procedures, and Rules for Evaluating Student Achievement at Various Levels

	Ministry	Board/District	School	Class	Student
<i>Policy</i>	The Ministry will provide a model of evaluation to be used at all levels in the system.	The Board will determine the degree to which its students are learning skills valuable to themselves and their society.	The school will provide regular reporting to the Board and to parents on the achievement of its students.	Each teacher will maintain complete and up-to-date records on all students in the class.	Counsellors will arrange for special needs so that the academic potential of each student can be realized.
<i>Procedures</i>	The Ministry will finance and coordinate the development of a pool of instruments for use with all grades and programs for the above purposes.	The Superintendent of Curriculum will develop a plan whereby Language Arts, Mathematics, Social Studies and Physical Education achievement will be assessed every three years.	Each department head will monitor the adequacy and accuracy of the achievement information produced in the department.	A variety of evaluation instruments will be used by teachers on a continuous basis.	Early identification of need will be instituted in the first two months of each school year and where needed thereafter.
<i>Rules</i>	The Ministry will solicit materials for the pool from many levels of education in the province.	All principals will be responsible for compiling school-wide data on areas being evaluated and for ensuring the achievement data are collected reliably.	Standard formats for grade reporting will be followed by all departments.	Students will be informed of the results of all evaluations conducted on them.	Recommendations for special needs assessments will be approved by the classroom teacher in consultation with special counsellors.

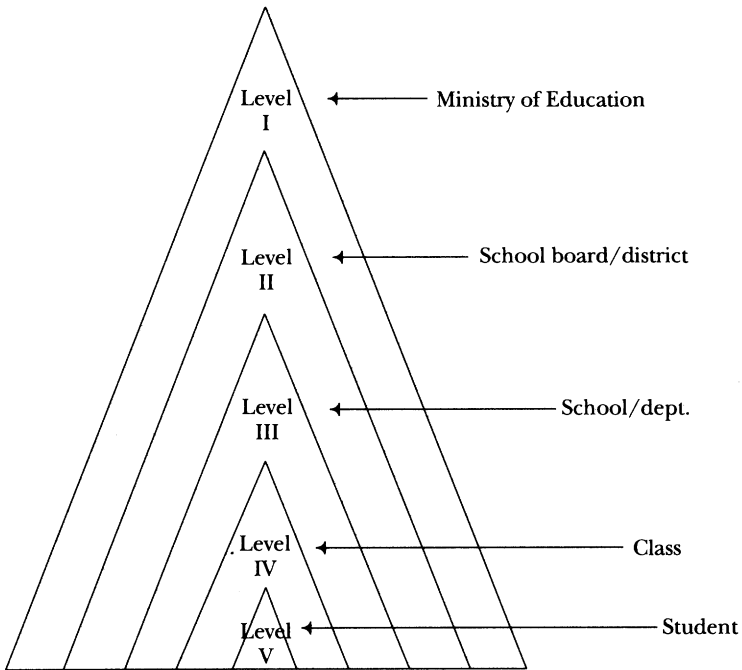


FIGURE 1

An Hierarchical View of a Provincial Educational System

setting for the study of educational evaluation in context, an ecology of evaluation.

MODELS OF THE EDUCATIONAL SYSTEM

A more exact view of the educational system from the provincial level on down is provided in Figure 1. First of all, Figure 1 depicts the educational system as a triangle, as Anthony (1965) proposed, reinforcing the perspective that the base of the triangle, the operational core, is the location of most of the activity of any viable organization (Emery & Trist, 1969). Secondly, the concept of the hierarchical yet interdependent levels is included by showing the nested levels of responsibility and authority in the overall educational system.

This model could easily be expanded or contracted to illustrate the chain of command, the variety of hierarchical levels that could be implicated in different provincial evaluation policy statements. The diagram also usefully shows different levels of policy, from complex policies originating outside an

institution to simpler ones affecting one and only one level in an organization. Overall, then, Figure 1 demonstrates a flexibility and verisimilitude lacking in the more discrete cellular representation offered in Table 2. Figure 1 is preferable not only because it provides a means for documenting policy, but also because it has the capacity to pinpoint not only which levels might be interacting in a particular policy, but also the dependency or interrelationship between these different levels.

Despite these merits, Figure 1 does not show clearly the stages included in policy enactment. That is, how are these interdependent levels involved in a particular policy? And, how is the policy initiation stage connected with the policy enactment stage? Parsons (1956) has indirectly answered both these questions with his conceptualization of a social system. He proposed that organizations as systems required three interdependent subsystems—the strategic, administrative, and operational subsystems. While each subsystem carries out certain activities or functions in a relatively independent manner, collectively the subsystems require some overall coordination through an intermediary subsystem to attain a system's overarching goals. This middle subsystem is the "linchpin" referred to by Fullan and Park (1981). It provides interpretation of and support for the policy so that it can be carried out.

The first subsystem, the strategic subsystem, possesses both an external and an internal focus. Its external view is aimed at defending, legitimating, and maintaining the organization, its goals, and its domain in its environment to ensure survival and growth of the organization. Its internal focus deals primarily with policy formulation, in order to provide the internal direction necessary for overall organizational goal attainment. The second subsystem, labelled here the administrative subsystem, carries out all functions necessary to integrate those policy initiators with the policy implementors. Those activities are coordinative in nature, including monitoring and control aspects as well as the allocating and scheduling of resources provided by the strategic subsystem. The third, the operational subsystem, is concerned with policy implementation, using the resources assigned to transform policy intention into action.

The model building thus far had led to the conclusion that any policy, including that of evaluation, must be communicated by the leaders at the apex of an organization (those in the strategic subsystem) by way of the middle managers (those in the administrative subsystem) to front-line individuals (those involved directly in operations). Stoner (1978) and others have suggested that policy communication, from articulation to enactment, occurs in three distinct steps: (1) policy, duly broken down into (2) procedures, made even more specific in the form of (3) rules. These stages match functions in each of the Parsonian subsystems—the strategic subsystem formulates policy then transformed by the administrative subsystem into

coordinative procedures, in turn translated by the operational subsystem into even more detailed guidelines known as rules.

Acknowledgement of this format for devolution of policy in an organization allows for renaming of the three subsystems in the triangular-shaped system. First, the strategic subsystem formulates policy. Then the administrative subsystem interprets the policy into procedures, the standard ways and means to be followed. Finally, the operational subsystem rephrases the procedures into even more specific guidelines known as rules, the outline of the uses of the resources in order to implement policy.

To summarize, each of the five hierarchical but interdependent levels comprising a provincial educational system (Figure 1) can be considered as a system in itself. Furthermore, each level is comprised of three subsystems. What remains, then, is to combine these two ideas in such a way as to reveal their interrelationships. Given the variety of ways in which such a mapping could conceivably occur, which one would be the most productive and realistic?

Beer (1979), in his ongoing study of cybernetics as applied to open systems theory, grappled with a similar problem. He noted that any system becomes increasingly process-specialized or functionally differentiated over time. Rather than adding more discrete subsystems, Beer contended that the model must highlight the interactive nature of these systems. Although he believed each viable system was comprised of five, not the more commonly accepted three, subsystems, he argued that the embedding occurs in the lowest tier, the operational subsystem, and that within each operational subsystem nests another complete system. This embeddedness in operational subsystems could continue *ad infinitum*, depending on the number of levels where information is available and useful. This "systems within systems" concept is known as Beer's theory of recursion. It recognizes, as no other known model does, both the interdependence of systems *and* the interdependence of functions at all the hierarchical levels of decision making that occur in the implementation stage of any process.

Parsons' model, with the incorporation of a concatenated version of Beer's conceptualization of a system (using three rather than five nested subsystems), can accurately represent any system. Figure 2 illustrates Beer's theory of recursive systems, but using only two systems in order to demonstrate where (in which of the three subsystems) the nesting occurs. No doubt Beer would contend that Figure 2 portrays a system more complex than Parson's because Beer depicts how one complete system "recurs" within one part of a larger system. This adapted version of Beer's model (Figure 2) was developed by Rees (1983). In a study using this model, she demonstrated its usefulness for describing the hierarchical yet integrated nature of manpower training at the federal, provincial, and local levels.

The final diagram, Figure 3, combines the concepts defined previously to

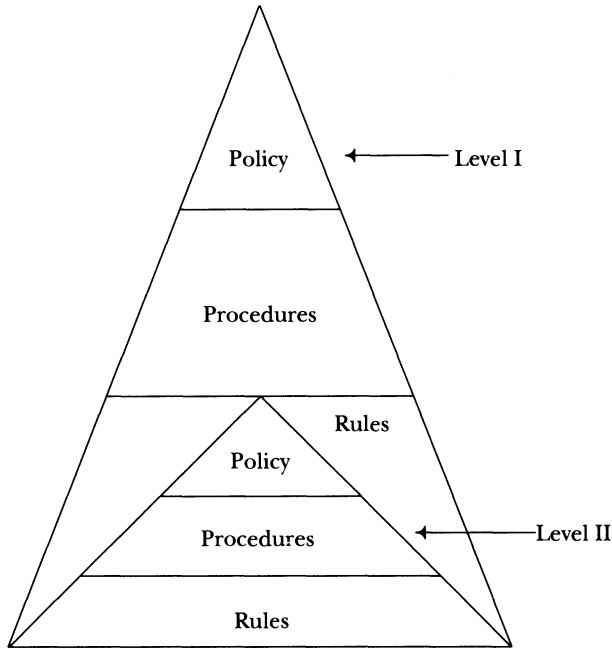


FIGURE 2

A System With One Embedded Level (following Beer, 1979)

illustrate how an educational policy might be captured. It seeks to delineate the key assumptions made thus far:

1. a provincial educational system is comprised of (at least) five hierarchical levels;
2. each of these levels (except the top one) is nested within another level through that latter system's operational subsystem;
3. each of these levels is viewed as a system containing three functionally differentiated but interdependent subsystems known as the strategic, administrative, and operational subsystems;
4. each of these levels can be considered separately for certain purposes only. For other purposes, each must be seen interactively, with different levels of responsibility and authority occurring between levels as well as within levels.

AN ILLUSTRATION

It remains now to incorporate the evaluation model in this administrative structure. Table 2 provided a description of certain possible policies, procedures, and rules for the evaluation of student achievement. No attribution to

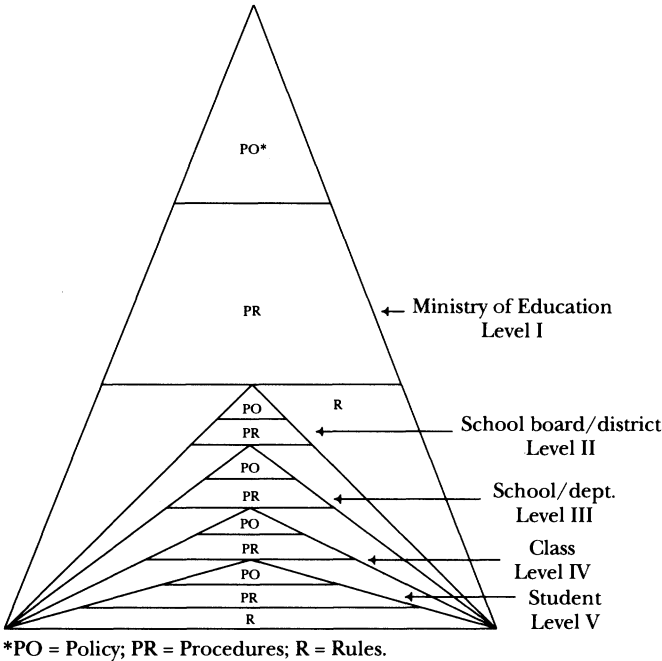


FIGURE 3

A Provincial Educational System With Four Levels of Recursion

any particular province, board/district, or school should be made even though most of these statements were adapted from actual documents.

The recursive and embedded qualities of the educational framework are evident at several places. The rules statement from the Ministry level, for example, becomes the key element for the Board's policy and its implementation. The school's policy statement on regular reporting results in a particular procedure for special needs' assessment by counsellors, and so on. Obviously, only a complete elaboration of all the policies, procedures, and rules at all levels would allow a full description of the interactions to appear. We are conducting such a project for two provincial systems.

The application of the model of evaluation to this environment raises involve the following questions: Are these instruments, once developed, reliable enough to support the various judgements that are being advocated within the many levels of the system? Would these instruments validly support the types of judgements called for in the policy? Would decisions made partly from these judgements be useful in furthering the educational enterprise in the school, the Board, or the province? To what extent were alternatives considered and rejected? The answers to such questions would go far in

establishing the efficacy of any policy concerning evaluation in the context in which it must operate.

CONCLUSION

This paper has proposed a theoretical model of evaluation that would fit within an equally theoretical conception of the structure and functioning of a complex educational enterprise, a provincial system of education with its interrelated levels. The major assumption of the paper is that attempts to conceive of evaluation at whatever level, for whatever purpose, that do not locate themselves in this institutional context are doomed to inadequacy. Conversely, a model integrated into this context might sensibly describe critical aspects of evaluation as they occur or should occur within any educational setting.

REFERENCES

- Anthony, R.N. (1965). *Planning and control systems*. Boston: Harvard School of Business Administration.
- Beer, S. (1979). *The heart of enterprise*. Chichester, UK: Wiley.
- Cronbach, L.J. (1971). Test validation. In R.L. Thorndike (Ed.), *Educational Measurement* (2nd ed.) (pp. 443–507). Washington, DC: American Council on Education.
- Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: John Wiley & Sons.
- Emery, F.E., & Trist, E.L. (1969). Socio-technical systems. In F.E. Emery (Ed.), *Systems thinking* (pp. 281–296). New York: Penguin.
- Fullan, M., & Park, P. (1981). *Curriculum implementation*. Toronto: Ministry of Education, Government of Ontario.
- Leithwood, K.A., Wilson, R.J. & Marshall, A.R. (1978). *Final report: The development and implementation of a strategy for comprehensive program evaluation for Ontario school systems*. Toronto: Ministry of Education. Unpublished.
- Leithwood, K.A., Wilson, R., & Marshall, A.R. (1981). Increasing the influence of evaluation studies on program decision-making. In A. Lewy & D. Nevo (Eds.), *Evaluation roles in education*. London: Gordon & Breach.
- McLean, L.D. (1982). *Report of the 1981 field trials in English and mathematics: Intermediate division*. Toronto: Ministry of Education.
- Messick, S. (1975). The standard problem: meaning and values in measurement and evaluation. *American Psychologist*, 30, 955–965.
- Parsons, T. (1956). Suggestions for a sociological approach to the theory of organization. *Administrative Science Quarterly*, 1, 63–85, 225–239.
- Rees, R. (1983). *The interorganizational collectivity: A study of the Manpower institutional training system in Manitoba*. Unpublished doctoral dissertation, University of Toronto.
- Scriven, M. (1967). Methodology of evaluation. In R. Tyler, R. Gagné, & M. Scriven (Eds.), *Perspectives of curriculum evaluation* (pp. 39–83). Chicago: Rand McNally.

- Scriven, M. (1974). The evaluation of educational goals, instructional procedures, and outcomes *or* The iceman cometh. In J. Blaney, I. Housego, & G. McIntosh, (Eds.), *Program development in education* (pp. 134–162). Vancouver: University of British Columbia.
- Stoner, J.A.F. (1978). *Management*. Englewood Cliffs, NJ: Prentice-Hall.
- Stufflebeam, D.L., Foley, W.J., Gephart, W.J., Guba, E.G., Hammond, R.L., Merriman, H.O., & Provus, M.M. (1971). *Educational evaluation and decision-making*. Itasca, IL: F.E. Peacock.
- Weick, K.E. (1977). *The social psychology of organizations* (2nd ed.). Reading, MA: Addison-Wesley.
- Wilson, R.J. (1989). Evaluating student achievement in an Ontario high school. *Alberta Journal of Educational Research*, 35, 134–144.

Robert J. Wilson and Ruth Rees are in the Faculty of Education, Queen's University, Kingston, Ontario, K7L 3N6.