Practical Foundations for a Science of Education

Presented, to the scientific and education community on 18 July 2019, at the 15th International History, Philosophy and Science Teaching Conference (IHPST2019), Thessaloniki, Greece

Paul Zachos
Association for the Cooperative Advancement of Science and Education
paz@acase.org
1-518-583-4645

Monica De Tuya
Association for the Cooperative Advancement of Science and Education
monica@acase.org

A Science of Education?

Nearly 100 years ago in his address to the International Honor Society in Education, John Dewey posed the question — "Can there be a science of education?" He elaborated: "Are the procedures and aims of education such that it is possible to reduce them to anything properly called a science?" He followed his own questions with a conclusion — that the quality of being scientific is not inherent in the subject matter but rather in the methods by which the subject matter is approached. There would be no reason then to exclude education from the objects that could be studied scientifically. Dewey then laid out the issues that needed to be considered in order to properly address these questions, specifying the qualities "…in virtue of which various fields are called scientific" (Dewey, 2013, orig. 1929).

Among these qualities are:

"…an intellectual technique by which discovery and organization of material go on cumulatively, and by means of which one inquirer can reappear the researchers of another, confirm or discredit them, and add still more to the capital stock of knowledge" (ibid. p 9).

Dewey expressed the expectation that a science of education would offer a number of benefits including a deepening understanding of facts; a path towards discovery of new facts; the ability to confirm; and, anticipating Karl Popper's perspective, the ability to disconfirm theories and findings (Popper, 1959). Dewey saw the scientific approach as providing, at the same time, both a sense of control *and* liberation from dull routine and blind tradition, and an understanding that would persistently bring a new eye to phenomena.

Overall, he imagined, "…the ways by means of which the functions of education in all its branches and phases—selection of material for the curriculum, methods of instruction and discipline, organization and administration of schools—can be conducted with systematic increase of control and understanding…" (Dewey, ibid).

Two Problems Facing the Establishment of a Science of Education

In the ensuing century vast resources and effort were invested world-wide in building an educational research community devoted to realizing the goals of bringing the benefits of science to the field of education. Research has been conducted under the auspices of governments, universities and private foundations. Research publications flourish. The breadth of this initiative's accomplishments is presented in *Educational Research: A Century of Discovery* (Alexander, Levine & Tate, 2016.) We might well declare this to be the first great age of educational science were it not for the fact that a number of serious problems persist.

One problem was identified in a special issue of *Educational Measurement: Issues and Practices* on "Changing the Way Measurement Theorists Think about Classroom Assessment" (Brookhart, 2003). The theme of Brookhart's article, and that of the other prominent educational researchers who contributed to that issue, was that the psychometric paradigm, which has dominated the field of educational measurement since the beginning of the 20th century, was not producing information that was useful and practical for classroom teachers. Of particular interest to us here are the articles by Pamela Moss (2003) and Jeffrey Smith (2003) which detail how the criteria of validity and reliability that have undergirded social science research throughout that century have not been made relevant to classroom assessment.

At the end of that century Mauritz Johnson, in his last published article, highlighted a second problem. Johnson lamented the fact that a basis for cumulative knowledge, of the kind that Dewey hoped for, had not yet been established (Johnson, 1985).

Our attention will be focused on these two problems as they derive from a common underlying source which remains the primary impediment to establishing a scientific basis for the study of educational aims and procedures. The problem can be summarized briefly as follows: *The field of education has been working with the wrong information and misusing the information that it has.* This follows from a careful look at what Dewey called the "subject-matter" of education.

<u>What is the 'Subject Matter' of Education?</u>

Johnson's contribution to the field of education stands in notable comparison to one of Isaac Newton's less appreciated contributions to the science of physics—

> "…the new discipline of physics could not proceed until Isaac Newton appropriated words that were ancient and vague — force, mass, motion, and even time — and gave them new meanings. Newton made these terms into quantities, suitable for use in mathematical formulas". (Gleick, 2011, p.7)

In attempting to parse the confusing rhetoric of educational discourse Johnson encountered dramatic inconsistencies in use of the term *curriculum* This led him to discover the central consideration in characterizing or defining any educational enterprise — identifying the human capabilities that are to be developed (Johnson, 1967). He realized that the critical unit of thinking and action for educational processes must be the *intended learning outcome* (Johnson, 1977). On this basis he was able, with clarity and precision, to distinguish *curriculum* from *instruction*, that is to distinguish what is to be learned from the efforts made to facilitate that learning. This distinction can be found as early as the 17th century (Comenius, 1953, orig. 1645). It is operational in the work of Ralph Tyler (1949) and was incorporated into his later accomplishments in building the National Assessment of Educational Progress.

Building on Johnson's discovery, Zachos and Doane (2017) were able to apply the notion of intended learning outcomes (ILOs) to bring precision to thinking about two other essential educational activities — *assessment* and *evaluation.* Educational assessment was designated as the activity of finding out how well ILOs have been attained. Educational evaluation could then be productively re-construed as the use of assessment information (and other relevant information) in making decisions about how to increase the value of curriculum, instruction, assessment, and evaluation itself! Here was a precise delineation of the aims and procedures of education, a practical basis for bringing Dewey's "functions of education in all its branches and phases" (Dewey, ibid) into a coherent framework.

In summary, once education has been clearly recognized to be any enterprise devoted to the realization of intended learning outcomes, *curriculum* can be unambiguously identified as the targeted ILOs themselves, *instruction* as the direct efforts made to help students attain the ILOs, and *assessment* as the measurement of degree of attainment of ILOs. Assessment then can be clearly recognized as the scientific wing of the educational enterprise, the means of generating systematic knowledge on how well ILOs have been attained. Interestingly, once the *subject matter* of education, its aims and procedures, have been adequately delineated it becomes evident that many activities that take place in educational institutions are not really educational at all, i.e. they do not have to do with helping students attain ILOs. Activities such as taking attendance, physical exercise, and even conventional testing and grading can be seen to be essentially non-educational in their nature.

<u>What is educational information?</u>

Indeed, it becomes clear that conventional test scores and grades do not provide *educational* information at all because the information they do provide cannot be used to support teaching and learning. The reason for this is that conventional test scores and grades are generated by combining the results of assessments of student attainment on multiple distinct and diverse ILOs. This process of aggregating scores from distinct ILOs masks the information concerning how well each of the distinct capabilities was attained. But it is information on how well the discrete ILOs have been attained that is the only real basis for planning, carrying out and evaluating instruction. The scores obtained from aggregation across discrete ILOs obscure the information needed for educational decision making. Test scores composed of aggregations of performance on different learning goals permit only comparison of relative performance of test takers and say nothing concerning what was and was not learned. This is the underlying source of the problems associated with high-stakes, norm- referenced testing and grading. Instead of providing an informational basis for making sound educational decisions they can serve only as a tool for social management based on level of success in test performance. *The field of education has been working with the wrong information and misusing the information that it has.*

Truly educational information then is information about the level of attainment of discrete ILOs at a level of specificity appropriate for instructional decision making. We refer to this level as the *critical level of* specificity and ILOs that have the properties of discreteness and practical level of granularity as *practical learning* goals. The process of educational assessment is then productively construed as the collection, analysis and presentation of information on how well practical learning goals have been attained. The information resulting from assessment and aligned to those practical learning goals can be referred to as *practical learning* outcomes. The relationship between practical learning goals and practical learning outcomes permits operationalization of intended learning. From this, we now have a basis for cumulative knowledge; and a basis for unpacking the second problem identified, that of practical reliability and validity.

It is notable that one of the most substantial contributions to assessment of human capabilities in the 20^{th} century was Barbel Inhelder and Jean Piaget's empirical studies of the development of logic-mathematical thinking (1958) in children's approaches to scientific problems. The information that they collected and analyzed maps easily onto practical learning goals and outcomes, i.e. onto a teachable level of specification of capabilities (Zachos, Hick, Doane, & Sargent, 2000). Inhelder and Piaget had demonstrated a model for research and analysis of educational information that could actually be carried out by teachers in their own settings working with their own ILOs and scales of attainment.

The Quality of Educational Information

The special 2003 issue of *Educational Measurement* referred to earlier identified the failure of the psychometric paradigm to produce indicators of validity and reliability that would be practical and of use to teachers. However, it did not properly identify the cause. The true obstacle to practicality and usefulness is the same one that prevents conventional assessments from being educationally useful at all. This is that the resulting scores are aggregates of multiple distinct ILOs. This aggregation obscures the information needed for educational decision-making because it permits only comparison of relative performance of test takers and says nothing concerning what was and was not learned. The proper content of educational information is how well discrete ILOs have been attained. This will turn out to be the key to tackling the issues of validity and reliability in a practical way as well.

One of the signatures of the development of a scientific discipline is its success in generating productive and reliable technologies to support its investigations. Information derived from educational assessments, like any systematic measurement, must meet two scientific criteria — validity and reliability. Validity has to do with how well evidence supports the interpretations of assessment results in alignment with what the assessment is intended to measure (American Educational Research Association et al., 2014). Reliability refers to the consistency of measurement (Traub & Rowley, 1991). Since Smith's positing of the reliability problem for classroom assessment in 2003 (see above) substantial progress has not been made in developing a way to conceive of reliability that is practical and useful in actual educational settings. The reason for this is straightforward. The existing conceptualizations of reliability are tied to the psychometric paradigm and tests to that are built of composite measures which aggregate disparate learning goals. Measures of validity and reliability of composite test scores are of no more use in educational decision making than are those scores themselves. The psychometric indicators for validity and reliability are neither interpretable nor practical in instructional settings. However, working at the level of practical learning goals and outcomes can provide access to a simple way of studying consistency in judgments concerning students' levels of attainment. This method is referred to as *Inter-Rater Reliability (Cizek, Schmid, & Germuth, 2016)* (Cizek, Schmid, Kosh, & Germuth, 2016).

Imagine that a science teacher is constructing an activity that she hopes to use every year to assess how well her students have attained 6 ILOs related to grasping the phenomenon of floating and sinking. She administers the assessment to a group of students and invites two colleagues to review the results with her. She has created rules for judging students' levels of attainment and asks her colleagues to rate student performance on each of the ILOs. The ratings, and any comments the teachers have, are collected and organized using an online information system[1]. This scenario actually occurred, with education researchers acting as the science teachers. Figure 1 is a report from the system that highlights the results from this work, that is, agreement in the researchers' ratings of attainment on that ILO. What turns out to be of greatest value is the disagreements between raters. Why were there disagreements? Were the scoring rules inadequate? Were the scales of attainment ambiguous or poorly constructed? The answers to these questions can lead to refinement in the assessment activity. But more is possible. Perhaps there is an inadequacy in the learning goals themselves and they need to be re-conceptualized in some way. This points to a refinement of the curriculum. It also often turns out that one or more participants in such an activity are lacking an understanding in the underlying subject matter, possibly even the convening teacher herself. Professional development is warranted. The results of an educational scientific investigation of this sort, using the type of technology illustrated, can lead to numerous discoveries and practical refinements in an education program and can result in the building of community around helping students to attain ILOs, indeed a community of educational science! Through the study of inter-rater agreement and disagreement the scientific criterion of reliability can be brought to bear on the conduct and improvement of ordinary educational activities.

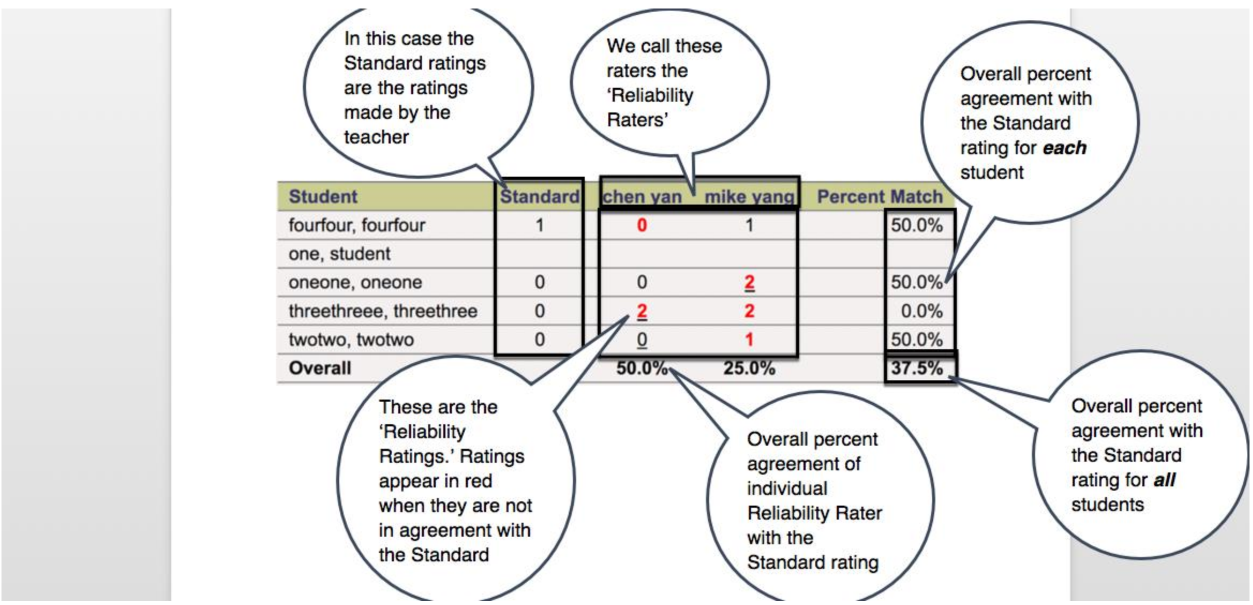

Figure 1. Results of an Inter-Rater Reliability Study for a Practical Learning Outcome

<u>A Community of Educator Scientists</u>

With a reliable information base comes the possibility of building cumulative knowledge. It is thanks to this intersection of scientific method, educational information and stakeholder engagement that Dewey's hope that "one inquirer can reappear the researchers of another" and "add still more to the capital stock of knowledge" can be realized in education. Implicit is the creation of a community of scientific practice within the educational enterprise itself. More than

[1] Visit ACASE's Online Assessment Information System at http://www.scientificinquiry.org/login.asp

simply formal obligations, assessment and evaluation can become a pragmatic social endeavor, inviting and encouraging discovery, collaboration, and communication among stakeholders in the educational sphere, indeed all those who are invested in the positive outcomes of educating well.

An Illustration of a Scientific Analysis of Educational Information Incorporating a *Core Capability*

Educational Assessment and Evaluation are often considered to be subsidiary rather than fundamental educational activities. Moreover, they can be expensive endeavors, consuming the time and resources of educational practitioners and institutions. The hesitancy to allocate resources to these activities is another reason why they have not taken their rightful place alongside curriculum and instruction as essential educational activities. For this reason, it is important to focus attention on assessing the attainment of the most critical capabilities first — the ones that make the greatest contribution to concurrent learning, subsequent learning and application to the world outside of the educational setting. We use the term *core capabilities* to refer to concepts, skills and dispositions that have these qualities of creating value.

Let us consider proportional reasoning as an example, a capability that is fundamental to the basic STEM disciplines as well as to the social sciences. Figure 2 highlights one student's attainment of the ability to **Apply Ratios and Proportions in Practical Problem Solving** (in this case it is the application of proportional reasoning to problems involving relative densities). The student's attainment is shown at three points in time each associated with an assessment of that capability. Three levels of attainment (0-2) are considered at each point in time. Performance of the class of students as a whole is displayed as a point of reference. This graphic demonstrates how cumulative knowledge of attainment of a practical learning outcome can be maintained and displayed to assist a teacher in planning and evaluating instruction. The teacher's judgment can be enhanced by estimates of the probability of the student actually being at the level of attainment indicated by the most recent assessment. Such probabilities can be generated using Bayesian algorithms that incorporate the history of information on attainment of that learning outcome, but also all other information in the student's records that might have predictive value concerning the level of attainment. Thus, the great accumulations of student information sitting in educational institutions can be put to productive use in a model of scientific data analysis that supports teacher decision making.

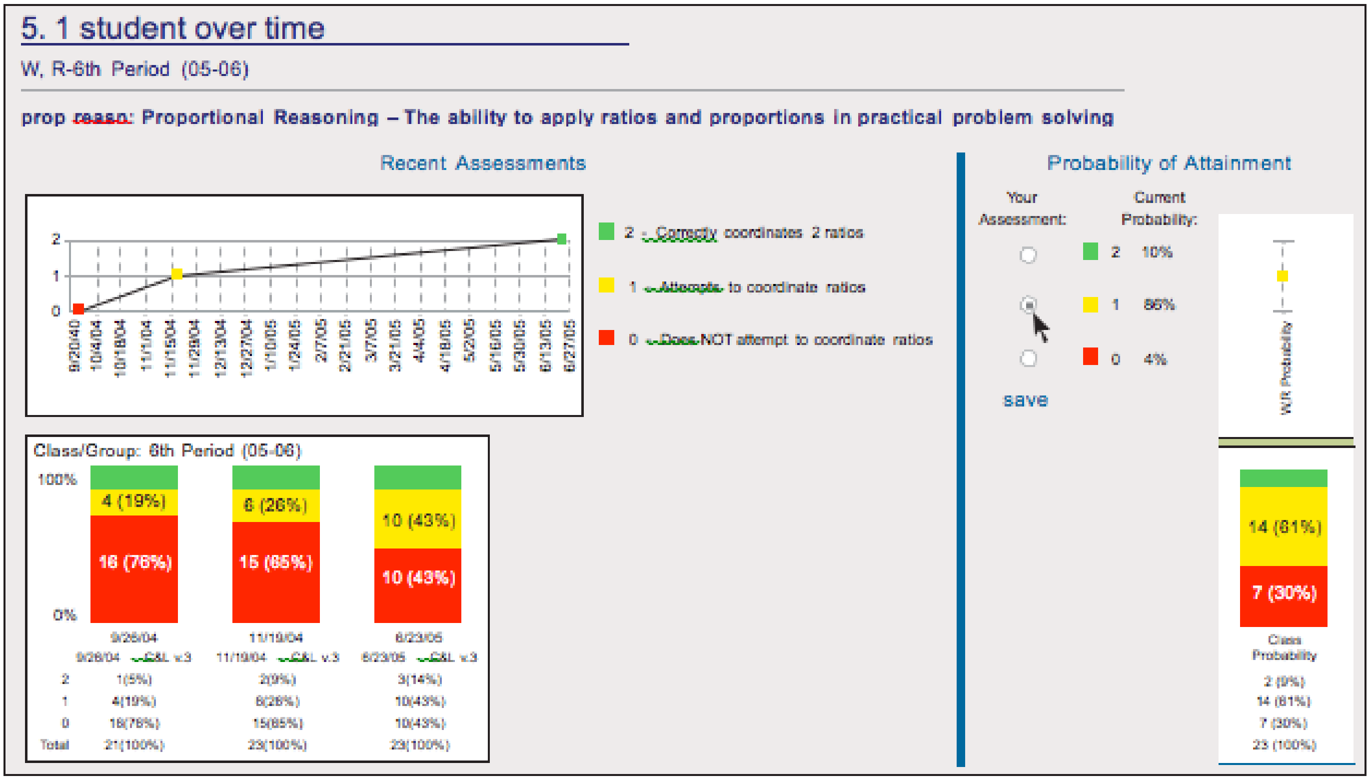

Figure 2. Cumulative Knowledge of Attainment Enhanced with Estimates of Probability (from *Knowing the Learner: A New Approach to Educational Information (Zachos & Doane, 2016)*

One of the most productive features of working with practical learning outcomes is that the information needed by teachers for their decision making turns out to be precisely the same information needed by students, parents, supervisors, educational researchers, evaluation specialists, policy makers, indeed all stakeholders whose time and energy is devoted to helping students to attain the intended learning outcomes of an educational program.

Even further, Figure 3 shows that information on the attainment of practical learning outcomes can be aggregated in a productive way, across hierarchies of educational institutions to study the institutional contribution to attainment over time.

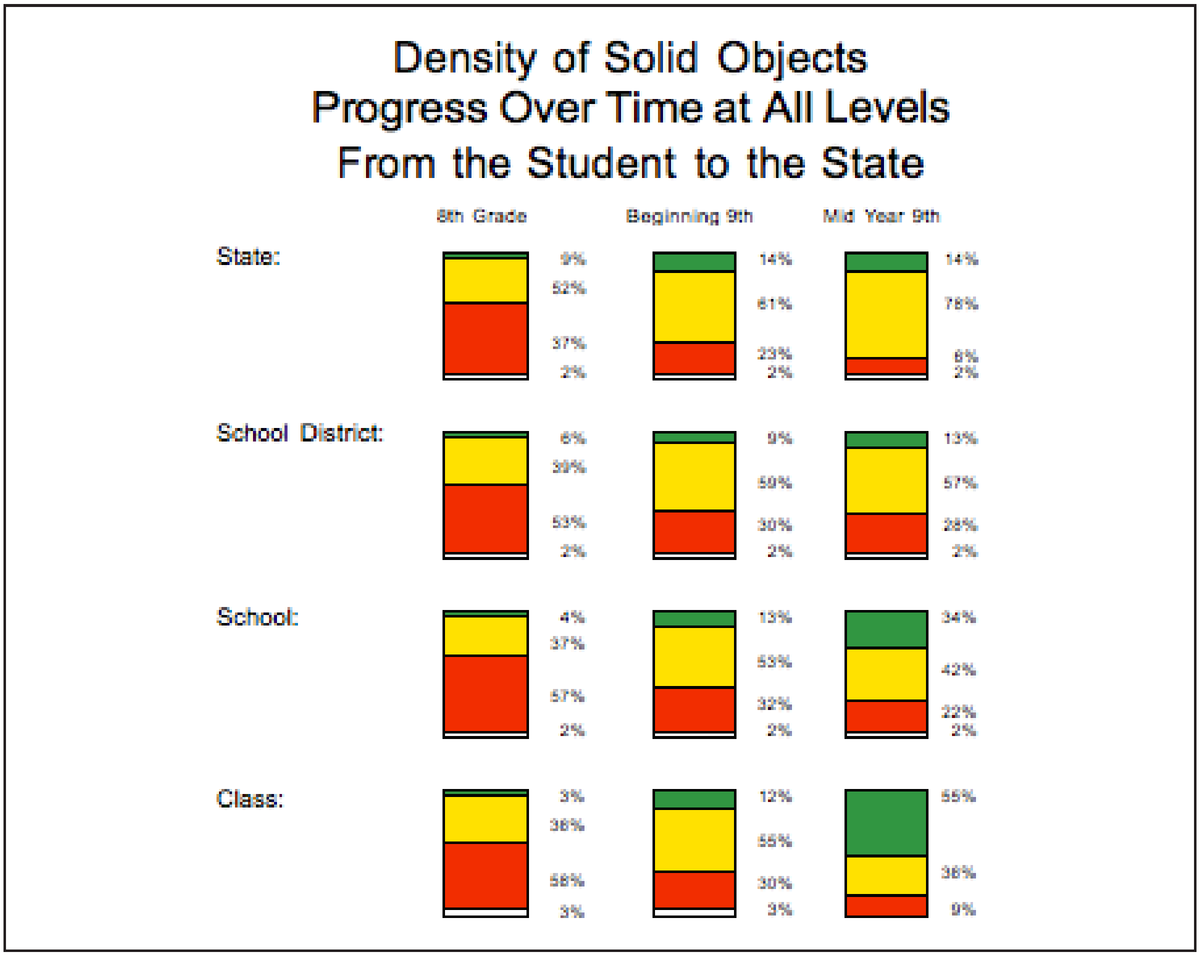

Figure 3. Aggregation of Information on Attainment of a Practical Learning Outcome across levels of educational institutions.

Reliable information on the level of attainment of discrete practical learning outcomes has the qualities needed to satisfy the most fundamental of Dewey and Johnson's hopes for a science of education: the ability to generate cumulative knowledge that can increase the ability of practitioners to understand and make intelligent decisions concerning teaching and learning.

<u>The Future of a Science of Education</u>

The hundred year stretch of educational research that we have been discussing has been centered on establishing valid scientific findings that could improve the provision of educational services to targeted populations. We have pointed out some of the shortcoming associated with historic attempts to reach that goal. But there is another perspective and possibility to be considered. Historically educational research has been devoted to seeking generalizable findings and best practices that can carried from one learning environment to another. To begin with, we maintain that establishing such findings and generalizations will necessarily depend on research being conducted at the level of practical learning goals and their outcomes as we have defined them for the reasons given. This has not been the case up until now.

But once information with such qualities becomes the basis of practice and research, a new perspective opens for educational science. Every piece of information on how well practical learning goals have been attained contributes to deepening knowledge of individual learners and how to serve them, what they need in the way of instruction, and what they are ready for. This actually constitutes a reversal of attention from interest in generalizations that are relevant to groups and turns it to obtaining richer knowledge concerning individuals. Such a reversal would constitute a paradigm shift (Kuhn, 1970) for the human sciences— a turn from generalization of findings to groups, in the direction of a deeper focus on the individual — leading to deeper understanding of the potentials for individual realization and its unique value to the world.

Principles on which a Science of Education can be Founded

This presentation has been, from the beginning, an explication of foundational principles for a science of education and a demonstration of their usefulness in educational practice. We conclude with a formal statement of the principles that underlie our work as educational scientists. They are, at the same time, intended learning outcomes. These are in fact the practical learning goals that we use as a basis for workshops we provide in the fundamentals of educational assessment and evaluation. We look forward to testing their efficacy, efficiency, and practicality in upcoming research and development projects. For simplicity of expression we use the term *learning goal* as synonymous with Johnson's *intended learning outcome*.

---

1. Using the concept of *discrete (un-aggregated) learning goals* to identify the end towards which all educational processes/activities are directed

2. Using the concept of the learning goal to distinguish *assessment* from the other essential features of educational processes/activities

3. Using the concept of the learning goal to distinguish *instruction* from the other essential features of educational processes/activities

4. Applying the concept of the *critical level of specificity* to educational planning and decision-making, i.e., distinguishes broader learning goals from those specified at a level appropriate for lessons and units of instruction

5. Applying the concept of *core capabilities* to educational planning and decision- making

6. Applying the distinction between *domains of learning goal*s to educational planning and decision-making

7. Distinguishing educational *assessment* from *testing* and *grading*

8. Distinguishing educational *assessment* from educational *evaluation*

9. Bringing concern for the notion of *validity* to challenges and problems in assessment and evaluation

10. Bringing concern for the notion of *reliability* to challenges and problems in assessment and evaluation

11. Applying assessment information to evaluate educational activities

12. Applying assessment information to build educational community

Works Cited

American Association for the Advancement of Science (AAAS). (1993). *Benchmarks for science literacy*. New York: Oxford University Press.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for Education and Psychological Testing.

Alexander, P. A., Levine, F. J., & Tate, W. F. IV, Eds. (2016). Educational Research: A Century of Discovery. *Review of Educational Research*, *40*. AERA: Sage Publications.

Brookhart, S. M. (2003). Developing measurement theory for classroom assessment purposes and uses. *Educational Measurement: Issues and Practice*, *22*(4), 5–12.

Cizek, G. J., Schmid, L. A., & Germuth, A. A. (2016). *A checklist for evaluating K-12 assessment programs*. Kalamazoo, MI: The Evaluation Center, Western Michigan University.

Comenius, A. (1953, orig. 1645). *The analytical didactic.* Chicago: The University of Chicago Press.

Dewey, J. (2013, orig. 1929) *The Sources of a Science of Education*. New York: Liveright.

Duckworth, A. L., & Yeager, D. S. (2015). Measurement Matters: Assessing Personal Qualities Other Than Cognitive Ability for Educational Purposes. Educational Researcher, *44*(4), 237-251.

Gleick, J. (2011). *The Information. A History. A Theory. A Flood.* New York: Pantheon Books, a division of Random House.

Inhelder, B., & Piaget, J. (1958). *The Growth of Logical Thinking. From Childhood to Adolescence.* (A. Parsons & S. Milgram, Trans.) New York: Basic Books, Inc.

Johnson, M. (1967). Definitions and Models in Curriculum Theory. *Educational Theory, 17,* 127-140.

Johnson, M. (1977). *Intentionality in Education.* Albany, NY: Center for Curriculum Research and Services, State University of New York at Albany.

Johnson, M. (1985). Curriculum's Missing Data Base. *Curriculum Inquiry, 15*(4), 359-360.

Kuhn, T. S. (1970). *The structure of scientific revolutions*. Second Edition. Chicago: The University of Chicago Press.

Moss, P. A. (2003). Reconceptualizing validity for classroom assessment. *Educational Measurement: Issues and Practice, 22*(4), 13–25.

Popper, K. (1959). *The Logic of Scientific Discovery*. New York: Basic Books

Smith, J. K. (2003). *Reconsidering reliability in classroom assessment and grading*. Educational Measurement: Issues and Practice, 22(4), 26–33

Traub, R. E., & Rowley, G. L. (1991). Understanding reliability. *Educational Measurement: Issues and Practice, 10*(1), 37–45.

Tyler, R. (1949). *Basic Principles of Curriculum and Instruction.* Chicago: University of Chicago Press.

Zachos, P. & Doane, W. E.J. (2017) *Knowing the Learner: A New Approach to Educational Information*. Vermont: Shires Press.

Zachos, P., Hick, T. L., Doane, W. E. J., & Sargent, C. (2000). Setting theoretical and empirical foundations for assessing scientific inquiry and discovery in educational programs. *The Journal of Research in Science Teaching, 37*(9), 938-962.