Open Access Library Journal
2021, Volume 8, e7628
ISSN Online: 2333-9721
ISSN Print: 2333-9705

Applying Kolmogorov's Proofs to the Evaluation of Instruction: Generalized and Particular (PART ONE)

John W. Oller

Institute for Pure and Applied Knowledge, Pittsburgh, USA
Email: john.oller@louisiana.edu

How to cite this paper: Oller, J.W. (2021) Applying Kolmogorov's Proofs to the Evaluation of Instruction: Generalized and Particular (PART ONE). *Open Access Library Journal*, **8**: e7628.
https://doi.org/10.4236/oalib.1107628

Received: June 10, 2021
Accepted: July 25, 2021
Published: July 28, 2021

Abstract

Successful sharing of information—positive (actual) knowledge about facts, skills that are imparted, abilities developed and expressed—is the implicit goal of instruction in all its varied forms. It is the goal of training athletes, dancers, and professionals in every walk of life from early childhood to the most advanced level of education. PART ONE introduces mathematical proofs showing that the interactional successes engineered by instructors, other things being equal, must trend toward 100% shared information—mastery of the course of study. Failed efforts trend toward a complete absence of shared information. All this holds independently for the subject-matter, methods of instruction, and the attributes conducive to instructional success. In Part One, the underlying proofs are united by a very simple proof from the theory of true narratives showing that every iota of knowledge that might be shared in any instructional context depends on the kind of representations found in true reports of actual experience. Empirical studies in Part One confirm the predicted agreement in diverse contexts on the elements of good teaching. In Part Two, Kolmogorov's proofs from 1933 are generalized, amplified, and tested empirically showing successful instruction converging toward 100% agreement on 1) subject-matter, 2) which methods of presentation and assessment work, and even on 3) the abstract criteria for successful instruction. At the same time, as the proofs also show, the cumulative effects of failed communicative efforts must and do trend toward zero shared information.

Subject Areas

Communication Theory, Education, Linguistics, Mathematics, Philosophy, Psychology

Keywords

Evaluation of Instruction, Kolmogorov Probability Theory, Peircean Logic of Relations, Polya's Central Limit Theorem, Pragmatic Information, Successful Communication, TNR-Theory

No person remains unchanged and has the same future efficiencies, who shares in situations made possible by communication [1].

1. Introduction

Not so long ago, I teamed up with some colleagues [2] to advocate the tearing down of disciplinary boundaries—to put gates in the walls and bridges across the gaps—that according to tradition are supposed to separate disciplines and subdisciplines in the departments, colleges, schools, and specializations within them all across the institutions of higher learning throughout the world. In addition to colleges and universities, here I also have in mind the training programs taking place in such contexts as international aviation, the business world, vocational schools, the information and technology sector, sports and the performance arts, and, in fact, in education in general. Today much of the teaching and training of interest is being done through the internet asynchronously—increasingly so since COVID-19. All this to say that here in both parts of this paper, I want to address instruction and training, educational interactions, in a general way, aiming to include them all. My purpose is inclusive, not because of any ambition like that of *Pinky and the Brain* [3] who wanted to "take over the world", but out of the desire to show how certain published mathematical proofs apply to all possible instructional contexts.

The kernel of truth to be developed here is implied in the opening quotation from John Dewey [1]. Communication enriches those who engage in it. It changes their efficiency toward an increasingly rich sharing of information much as a slight positive curvature, like the curvature of our planet, as Bernard Riemann showed [4] in 1873 (p. 10), ensures (if it exists) that the seemingly infinite expanse of space is both spherical and finite. Similarly, if there is even a modicum of shared knowledge between the interlocutors engaging in any form of instructional communication, the cumulative effect of their interactions over time must tend to produce a growing agreement. It must, as will be shown in Part One of this paper, and as will be demonstrated in greater depth and detail in Part Two, though progressing in a finite number of steps—and relatively few of them—proceed toward a theoretical limit of unity on three levels more or less simultaneously but quite independently. The three levels can be thought of as three distinct spheres where the outermost, largest, and last to be determined sphere contains the innermost, smallest, and first to be determined sphere, which is contained within the middle sphere which is determined after the first and before the third, and contains the former but is contained by the latter:

1) There must be increasing cumulative agreement on the subject-matter reflected in the scores on any form of valid assessment of the knowledge of the subject-matter of any course of study however it may have been presented.

2) There must also be increasing cumulative agreement on how and why distinct methods of interaction—explanations, demonstrations, directed practice, and so forth—serve to enhance the knowledge, skills, and abilities of the learners/students/trainees over time.

3) And, there must be increasing agreement on whatever abstract attributes of the instructor and the instruction that are required to achieve success in such communications.

Kolmogorov's proofs about probability, and generalizations of those proofs incorporating the theory of pragmatic information [5] [6] [7] [8] [9] as well as the theory of true narrative representations [10] [11], show that all three of the just mentioned levels of successful instruction tend to progress toward the theoretical limit of complete agreement, or we could say 100% unity, in what can be defined mathematically as shared information. In Part One of this paper besides illustrating key implications of the mathematical proofs for instructional communication, I provide empirical evidence from four intensive studies of the evaluation of instruction showing a very general concurrence of both teachers and learners about the attributes conducive to instructional success. Then, in Part Two, which is practically a self-contained presentation of the mathematical reasoning and its testable empirical consequences, additional experimental evidence is provided from multiple studies showing that tested competencies of students engaged with diverse methods of instruction and in different kinds of subject-matters trend toward the convergences predicted: these trends, as will be shown empirically in both parts of this paper, but especially in Part Two, do go as predicted—toward unity in successful instructional communication and toward zero shared information in random responses to test questions modeling completely failed instructional efforts.

2. Removing Barriers

The hypothetical "barrier" between the "classroom" and the "real world" was a figment. As soon as we take account of the fact that schools, industries, and businesses all seek to instill or enhance real and measurable "knowledge, skills, and abilities" (KSAs) [12] in recipients of instruction, the fictional separation of the instructional process from the "real world" vanishes. Also, as Kim noted in 2004, a more comprehensive look at the whole process of instruction from beginning to end, ensures the trend toward tasks that involve "real-life communication" (p. 1).

Realizing from the start that instructional success must be judged more or less in reverse—by asking what end result would that instruction enable if it were successful (see the research and methodologies being developed at Quality Matters [13])—our basic argument [2] from 17 years ago is sustained: we were already connecting the end of the road assessment procedures backward to the methods and subject-matter of the instruction. We noted back then that Eisner [14] had commented that "performance assessment [at the end of the road and along the way] is the most important development in evaluation since the invention of the short-answer test and its extensive use during World War I" (p. 2). The sort of assessment Eisner had in mind involved the end performances that students/trainees might be expected to be able to perform *after the planned course of study or training*. What, *at the end of the day*, is the instruction supposed to deliver in terms of KSAs? To show that I am not generalizing here in some unusual or grandiose manner, the Wikipedia article [12] on KSAs says:

> KSA statements are also known as Evaluation Factors. Other agencies sometimes call them "Rating Factors", "Quality Ranking Factors", "Knowledge, Abilities, Skills, and Other Characteristics", or "Job Elements".

Kim [15] defined "performance assessment"—based on five decades of prior research—as "any assessment procedure that involves either the observation of behavior in the real world or a simulation of a real-life activity with raters to evaluate the performance" (p. 1).

2.1. Training in an Exemplary High-Stakes Industry

Given that English is the required language for international aviation, our starting example in 2005 [2], was about getting air traffic controllers and pilots in international aviation fully up to speed in their English proficiency. The relevant research showed that tens of thousands of accidents [16], some of them horrendously injurious or fatal [16], were predominantly attributable to breakdowns in communication because the interlocutors were using a language not entirely intelligible to themselves [17] [18] [19] [20] [21]. The research showed that the high risk moments in international aviation arise when communications between pilots and air traffic controllers must take place in one or several distinct dialects of English between persons whose primary language is one of the other 7000 plus languages of the world [22].

2.2. Instruction Is Important

We supposed that the only adequate remedy for the well-documented breakdowns in communication in international aviation would be better English language instruction of the personnel—in particular air traffic controllers, pilots, and others involved in guiding passengers, cargos, and carriers. More specifically, we argued that the course of study/training should address the full range of contexts of communication about aircraft, airports, runways, equipment, traffic patterns, signals, lighting, and in general the subject-matter that interlocutors working in what may be a non-primary language for many of them, need to know at a level approximating perfection. The folks whose lives are on the line need to know that when the plane departs from a gate on one side of the world, for instance, that it is almost certain to arrive safely at the one it is destined to reach possibly on the other side.

Although not all contexts of instruction have the same high stakes as those in international aviation, many aspects of the problems presented in those contexts generalize appropriately. No matter what the stakes may be in any given area of instruction/training, the goal in general is to maximize success and minimize failure. No instructional process aims for failure any more than businesses are built with the intention of losing money rather than making a profit [23].

3. Three Critical Components of Instruction

To accomplish the goal of maximizing success (while minimizing the possibility of failure), it is generally agreed that 1) valid planning with respect to the subject-matter, 2) interesting and engaging communication of that subject-matter, and 3) valid assessment of the uptake of the subject-matter in real life scenarios aiming for optimal performance with multiple checks, backups, and follow-ups are essential. In fact, these three components are generalized, incorporated, and expressed in a multitude of ways in questionnaires for evaluating instruction, or most any training in higher education, the business world, sports, dance, and performance arts coaching, and in human industries in general. According to the recent work in Maryland by the growing community of researchers, teachers, and trainers following what is taking place at Quality Matters [13], the three principal components of any course of instruction must be brought into agreement (they call it "alignment") with each other.

So we ask, how can the policy-makers and the educators who implement the required procedures of instruction/training ensure with nearly complete confidence that the persons successfully working through some course of study in a given subject-matter will end up having the necessary KSAs for optimal performance at the end? In an industry such as international travel by air, the critical actors—pilots and air traffic controllers—need to be nearly perfect in their performances all the time in all the contexts that can and do arise. Expressing all this in the simplest language of Kolmogorov's probability theory and extensions of it to be made in the two parts of this paper, designers of instruction, it seems, must set up the instruction in such a manner as to make the likelihood of success as near to unity (perfect) as possible and the likelihood of failure as near to nothing (zero) as possible.

3.1. Questionnaires for the Evaluation of Instruction

The driving force behind the provision of instruction/training in any context is to enable recipient/participants to acquire KSAs shared by one or many persons who have either already achieved them or who know how to do so and are a little farther along the road toward doing so than their students are. The objective is always to increase the powers of students in the direction of those already possessed by the designers of the course of study—the teachers, coaches, trainers, and the like. Students hope to progress to the level of more advanced persons in the professions, businesses, industries, sports, or what-have-you who have already paid their dues and have achieved a desirable and advanced level in the KSAs of the "subject-matter". The instructors are not expected to be perfect performers but they need to be substantially beyond the level of the student/trainees who are invited, or possibly in some cases at universities and colleges, required to successfully complete certain courses of study. For this reason, almost universally, questionnaires for the evaluation of instruction/training, address the alignment/agreement of the measures or performances required at the end of the course of study with the subject-matter, the methods of teaching/training, and the overall course design used to instill the desired gains in the KSAs aimed at. The respondents may be asked: 1) to agree or disagree with a series of statements; 2) to select from multiple choices a level of agreement or disagreement, or to choose among several alternative answers the one that best expresses the evaluator's view; and/or 3) there may be open-ended questions asking the evaluator to compose a short response or an essay expressing something judged to be of importance about the course of study. Questions may ask what was liked or disliked, what should be kept or discarded, what worked and didn't work, etc. Here are some examples.

3.1.1. Evaluating Massive Open Online Courses

Cutting directly to the present tense of the evaluation of thousands of adult-level courses of instruction already reaching millions of people throughout the world, in 2021 Deng and Benckendorff [24] analyzed 8475 ratings and reviews from 1794 distinct Massive Open Online Courses (MOOCs) aimed, according to them at the "social sciences". They focused attention on positive reviews, which the mathematical proofs to be discussed below show are the proper ones of interest. If we are aiming for success in enhancing KSAs, the only instructional attributes of interest are the ones that enable progress toward success, and we are hardly interested (for reasons to be made entirely clear later in this paper) in all the different ways there may be to fail. Summarizing what has been proved mathematically by Kolmogorov and by generalizations of his proofs, the bottom-line is that there are always uncountably many more ways to fail—e.g., not to construct a lightbulb that works, not to succeed in a running a four-minute mile, not to get all the way to the top of Mount Everest, not to learn how to converse in Navajo, etc.—than there are ways to succeed. The proofs show why agreement on success must generally converge to unity (a probability of 1) precisely to the extent that the course has any success at all (like the slightest positive curvature of space assures us that it must be a finite sphere) while the likelihood of failure, amounting to a complete absence of communication must trend toward a probability of 0 information shared.

From their extensive computer-assisted analysis, Deng and Benckendorff distilled six critical components of successful MOOCs (loosely summarized in the "thematic map" of Figure 1): some of their criteria are generalizable to all courses of instruction aimed at advancing any learner's KSAs in any given subject-matter. I will sum these up in my own words:

1) realistic representation of the subject-matter,

2) requiring performance by students in real-life tasks demonstrating mastery of the subject-matter,

3) making the course experience interesting, engaging, so as to capture and hold the student's attention,

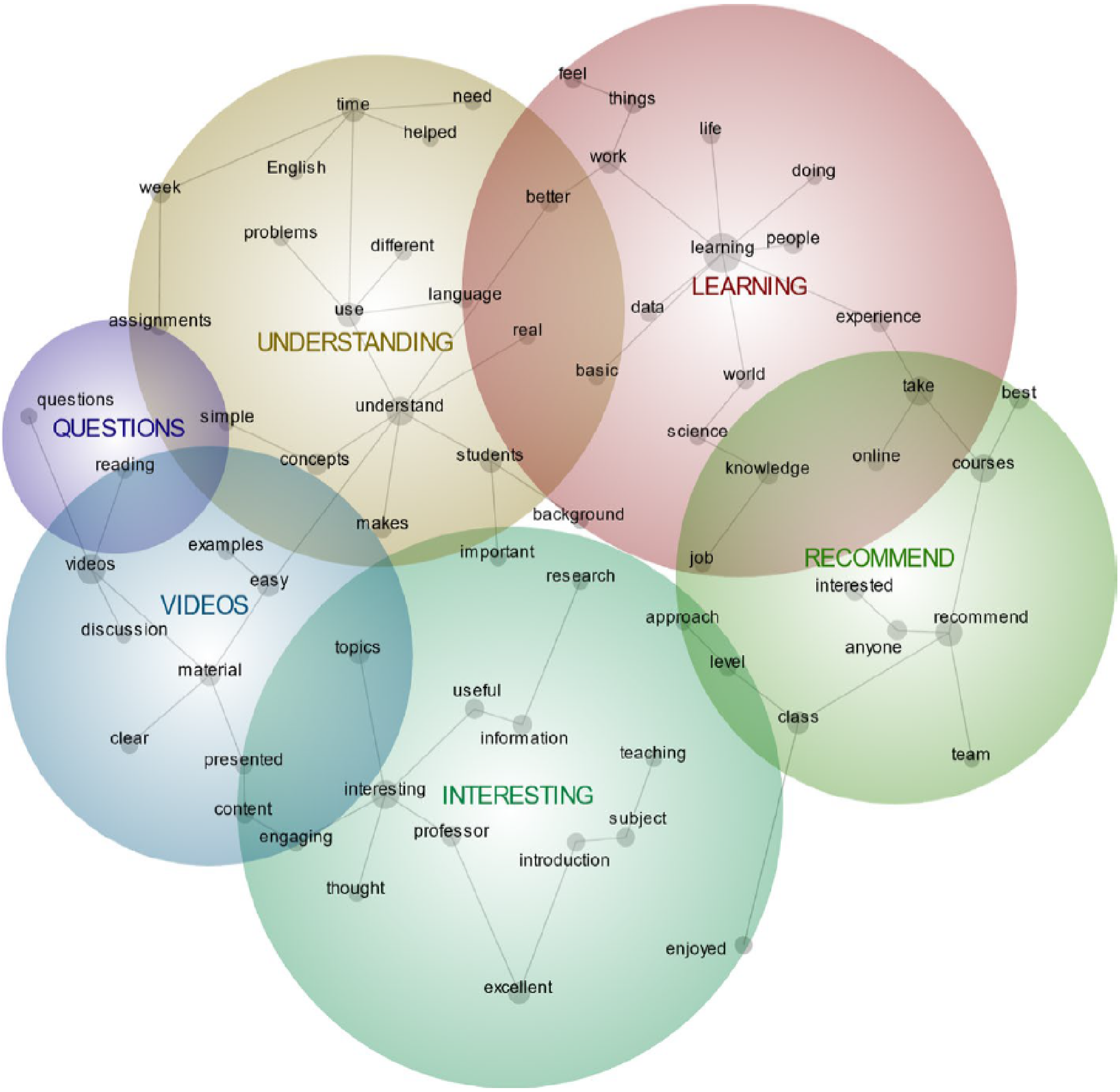


Figure 1. A "thematic map" from Deng and Benckendorff [24], "What are the key themes associated with the positive learning experience in MOOCs? An empirical investigation of learners' ratings and reviews", *International Journal of Educational Technology in Higher Education* (2021) 18: 9 at https://doi.org/10.1186/s41239-021-00244-3 reproduced by the authority of the Creative Commons Attribution 4.0 International License which can be viewed at http://creativecommons.org/licenses/by/4.0/.

4) investing sufficient resources and energy in the course design (video lectures for the MOOCs at issue in the study by Deng and Benckendorff) so that presentations are of high-quality,

5) illustrating difficult concepts and principles in intelligible and interesting ways in the presentations (video lectures, etc.),

6) inviting participation by incorporating comments and questions from students enabling them to become integral contributors helping instructors to make things clear to all the students.

One inevitable constraint on the computer-assisted analysis performed by Deng and Benckendorff is that their software performed its analysis mainly on the lexical items and phrases appearing in the questionnaires and responses of the hundreds of students included in the analysis summed up in **Figure 1**. Nevertheless, their analysis implicitly includes the three major types of agreement and at least suggests that even moderately successful instruction does converge toward 1) agreement on whatever the subject-matter is, 2) agreement on whatever methods of presentation successfully get the subject-matter across, and, finally, 3) agreement on the attributes conducive to instructional success in general.

3.1.2. "12 Amazing Course Evaluation Survey Questions"

Next we come to an organization online that identifies itself as the "Question-Pro" [25]. The website I am referring to at this link offers "Survey Software" among other instruction-related services, some of it for free, and some with differential pricing for collection and storage of more than 1,000 responses to any one of their constructed surveys. The authors of the site begin by telling visitors why colleges and universities "must… conduct evaluation surveys". They group their questions into three categories: those pertaining to the instructor, the course material and structure (methods) of presentation, and general satisfaction of the respondent with the course. Their "12 Amazing Course Evaluation Survey Questions" [25] consist of statements followed by an invitation to choose from a list of five choices expressing a relative frequency ranging from "almost always" to "almost never" or "strongly agree" to "strongly disagree". Here are the statements verbatim followed in brackets by [the type of scale used, and the evaluation target the authors of the website believe the item addresses]:

1) The instructor was well prepared for the class. [frequency, instructor]

2) The instructor showed an interest in helping students learn. [frequency, instructor]

3) I received useful feedback on my performance on tests, papers, etc. [agree-disagree, instructor]

4) The lectures, tests, and assignments complemented each other. [agree-disagree, instructor]

5) The instructional materials (*i.e.*, books, readings, handouts, study guides, lab manuals, multimedia, software) increased my knowledge and skills in the subject matter. [agree-disagree, course materials/methods]

6) The course was organized in a manner that helped me understand the un-

derlying concepts. [agree-disagree, course materials/methods]

7) The course gave me the confidence to do more advanced work in the subject. [agree-disagree, course materials/methods]

8) The examinations, projects measured my knowledge of the course material. [agree-disagree, course materials/methods]

9) I believe that what I'm being asked to learn in this course is important. [course materials/methods, agree-disagree]

10) I would highly recommend this course to other students. [agree-disagree, general satisfaction]

11) Overall, this course met my expectations for the quality of the course. [agree-disagree, general satisfaction]

12) The course was helpful in progress toward my degree. [agree-disagree, general satisfaction]

The authors then go on to mention different types of questions that may be included in course evaulation surveys and questionnaires. They differentiate "closed-ended" and "multiple choice" questions (which are really a single forced-choice category), from questions about preferences (which can be either forced-choice or open-ended questions eliciting short answers or longer essays about what students like or don't like), and, finally, they suggest questions in which respondents are asked to rank different options (possibly statements about the quality of instruction) in the order of their judged importance. Again, it is easy to see that the QuestionPro designers have in mind the three independent levels of agreement already introduced on subject-matter, successful methods of presentation, and the abstract qualities of instructional success.

3.1.3. Example from Laupper, Balzer, and Berger

For reasons that will become very clear later in Part One of this paper, the next study I want to examine closely compares the old-school paper and pencil surveys typically filled out in a classroom setting—or after the conclusion of a course of study to be returned to the college or training facility—with the contemporary online survey more widely used since COVID-19. The research to be reviewed from Laupper, Balzer, and Berger [26] is important for various reasons but among them is the possibility—if the methods should prove to be equivalent in their statistical properties—*of disposing of the claim that research with offline questionnaires is not relevant to the online type, and vice versa*. In fact, Laupper, Balzer, and Berger focused on the specialized agreement between obviously distinct methods of collecting data about the quality of instruction. Empirically, in **Figure 2**, and in their discussion of it, they have demonstrated precisely the sort of convergence that is logically required by the generalization of Kolmogorov's proofs, as discussed later, here in Part One of this paper.

Laupper, Balzer, and Berger [26] tested the convergence expected empirically and required mathematically, by applying a confirmatory factoring method. They collected responses to seventeen survey items from 232 respondents to an online version in 17 courses and from 231 similar respondents to an offline version in

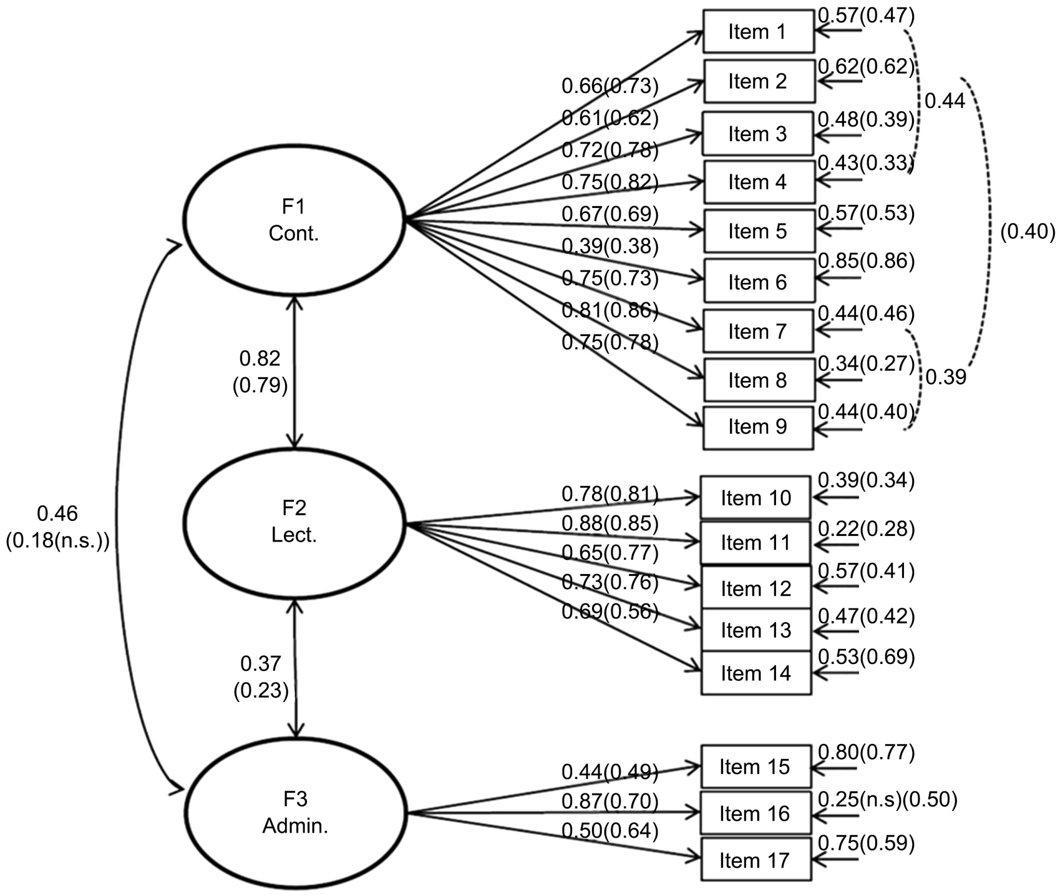


Figure 2. Confirmatory factoring of "Offline" and ("Online") evaluations for 33 vocational educational training courses sponsored by the Swiss Federal Institute for Vocational Education and Training Secretariat from the 17 items listed in the text. From Laupper, E., Balzer, L., & Berger, J.-L. (2020). "Online vs. offline course evaluation revisited: testing the invariance of a course evaluation questionnaire using a multigroup confirmatory factor analysis framework". *Educational Assessment Evaluation and Accountability*, 32(4), 481-498. Reproduced by the authority of the Creative Commons Attribution 4.0 International License which can be viewed at http://creativecommons.org/licenses/by/4.0/.

16 courses at a Swiss (German) institution of higher education. All 33 of the courses they examined were focused on the same sort of vocational and professional training. The specific items were these (in their own words):

1) The course content was adequately covered.

2) I consider that the impact of this course on widening my skills is low. (reversed)

3) The information provided is very useful to me.

4) I succeeded in meeting the objectives set in the course programme.

5) The documents provided proved helpful in understanding and learning.

6) During the course, there were a number of elements I did not understand. (reversed)

7) The methods used in the course were suited to the content.

8) I met my personal objectives for this course.

9) The course objectives were clear and understandable.

10) The lecturers were able to link theory to practice.

11) The lecturers explained the subjects in a clearly understandable way.

12) The lecturers seemed competent in their field.

13) The lecturers made the course lively and convincing.

14) Where there were several lecturers, their various interactions went smoothly.

15) I am satisfied with the course venue and infrastructure.

16) I am satisfied with the follow-up by the SFIVET [Swiss Federal Institute for Vocational Education and Training] Secretariat.

17) I am satisfied with the Click&Book registration process.

The confirmatory factor analysis summed up in Figure 2 shows that both the online and offline questionnaires resolved to three major components as expected by the designers of the questionnaire and which I paraphrase as follows:

1) Subject-matter addressed in… "nine items related to how adequate, informative, helpful, understandable, and goal-oriented" the course was judged to be;

2) methods of presentation by "lecturers" addressed in "five items" about "lecturers' competencies in linking theory with practice and creating a stimulating learning climate";

3) overall personal satisfaction with the course design, the evaluation process itself ("follow-up"), and the "registration process".

More importantly, the two methods of collecting data from students about their evaluations of instruction proved to be equivalent in all of the ways tested by Laupper and colleagues. Though superficially different in their response mode—in their deep structure from the point of view of the discursive meanings considered by intelligent human evaluators—the two modes of course evaluation appear to be equivalent for all practical purposes. Among other conclusions, it appears that research with paper and pencil, offline questionnaires and surveys of the quality of instruction in higher education, remains as relevant as it ever was in spite of the fact that educational institutions all over the world are proceeding more and more into the post-COVID-19 world of remote instruction via the internet. It appears that the "factor structure and relationships between the dimensions within it" as summed up in Figure 2 from Laupper, Balzer, and Berger are sufficiently similar across the two modes as to make them largely interchangeable. Also, those findings suggest that research into the evaluation of instruction relying on either of the two modes of eliciting responses from students should largely be generalizable to the other.

3.1.4. Quality Indicators Ranked by Persons of Increasing Experience

With all the foregoing in mind, but especially the findings from Laupper, Balzer, and Berger, I want to look back to an unpublished research project completed three decades ago. In the spring of 1991, I obtained permission from the University of New Mexico Institutional Research Board for a research project funded by the Public Service Company of New Mexico Foundation through their Distinguished Educator Award Program. It was titled: "Improving Instruction through On-Going Evaluation". At that time, the head of the University of New Mexico Committee on Teaching Enhancement, Charles Beckel, a Professor in the Department of Physics and Astronomy there, asserted that about 95% of UNM faculty see instructional quality as a low priority in terms of budget, rewards, etc., but more than 95% hold the quality of instruction to be one of their highest personal priorities[1].

Our charge was to address teaching quality at our institution from various angles. The committee eventually settled on administering a university-wide survey to give faculty and students a chance to rank the most common criteria for evaluating the success of university level courses as distilled from extant instruments for student evaluation of instruction in higher education. Members of the faculty were also invited to write an essay on what instructional excellence consists of and why it is a priority in their thinking as members of the university faculty. A cash prize was offered for the best essay as judged by the aforementioned subcommittee and plaques of recognition were to be offered for the 10 most interesting runners-up in the essay writing contest.

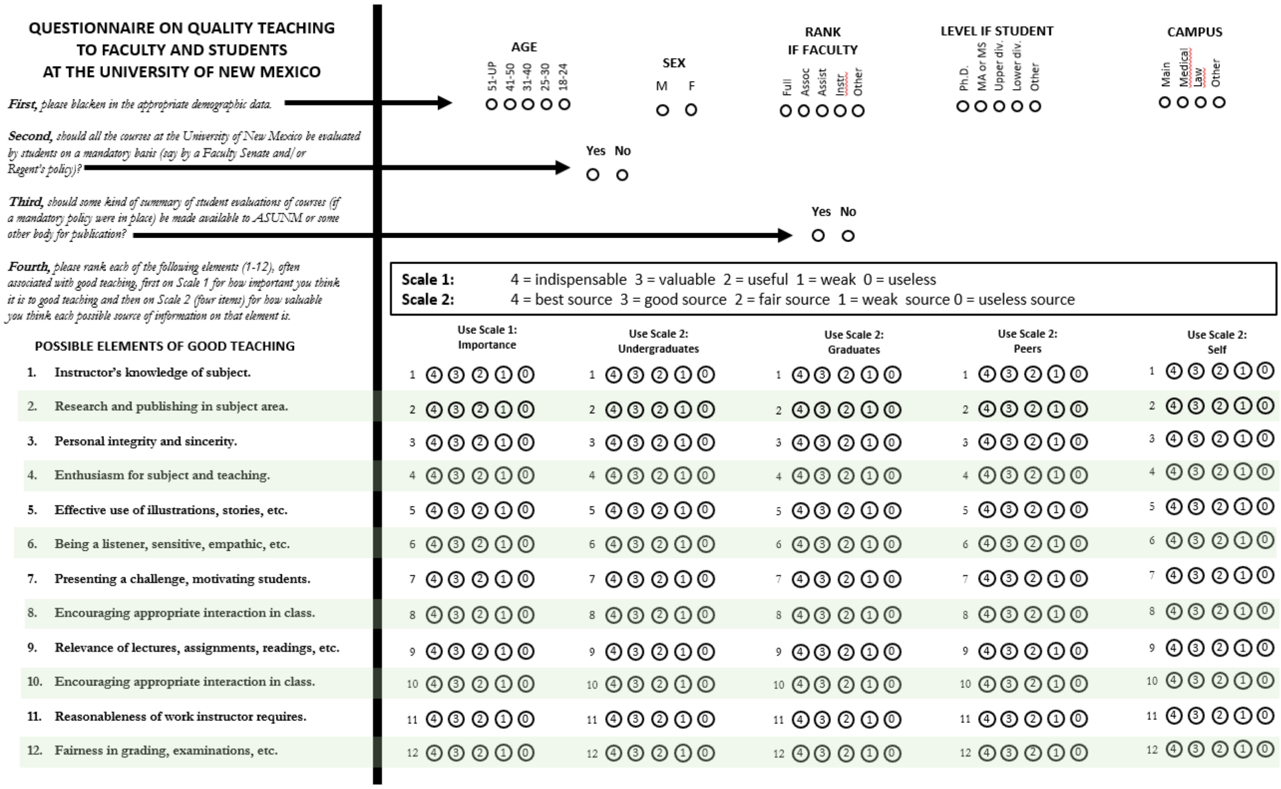
In the final phase of that research project, the questionnaire in **Figure 3** was distributed to faculty and students asking for age, sex, rank if a faculty member, and level if a student (**Table 1**), as well as a "yes" or "no" about whether course evaluations should be required and if so whether results should be published. Also faculty and students ranked the "possible elements of good teaching" and judged how well faculty peers, students at various levels, and the instructors, could rate those particular elements of successful instruction.

As seen in **Table 1**, 52% of the student respondents were undergraduates in the 18 - 24 age range, the bulk of them in their junior or senior year of undergraduate work, and 66% of them were on the main campus, with 24% at the Medical School and 2% at the Law School. The majority of faculty respondents were in the age range from 41 - 50 with 36% at 51 or older. No faculty respondents were in the 18 - 30 age range. There were almost as many Assistant Professors as Associate Professors, and only a handful of Instructors. For that reason, the calculations for the questions about mandatory evaluation and publication of the results were limited to Assistant, Associate, and Full Professors, though all student respondents were included in that portion of the table. Of the 327 faculty respondents who addressed our first question, 56% said "yes" to mandatory course evaluation and a slightly different group of 321 faculty members would favor publication of the results if such a mandatory course evaluation policy were in place. Students were even more positive in their responses to both questions with 86% of 848 respondents saying "yes" to both questions: students want courses evaluated and they want the results to be published. Also, there was an interesting trend of apparently increasing agreement across time for both faculty and students.

[1] At that time I was called on to head up a Subcommittee on Teaching Quality and Effectiveness consisting of Jean Civikly (Department of Communication), Chester Travelstead (Emeritus Professor and former Academic Vice President), Joseph Martinez (Head of Educational Foundations), Elsie Morosin (Professor in the Department of Nursing), John Saiki (School of Medicine), and Bill Hayward (Department of Health and Physical Education). At that time I was a Professor of Linguistics.

Table 1. Demographics and results on major questions in the survey on quality teaching at the University of New Mexico.

Age of Respondent	18 - 24	25 - 30	31 - 40	41 - 50	51-up
Student	52%	20%	16%	7%	4%
Faculty	0%	0%	23%	38%	36%
Faculty Rank	Other	Instructor	Assistant	Associate	Full
	6%	4%	26%	27%	35%
	Other	Lower Division Undergrad	Junior or Senior Undergrad	Masters Candidate	Ph.D. Candidate
Student Level	4%	17%	48%	21%	4%
Campus		Main Campus	Medical School	Law School	Other
Faculty		66%	24%	2%	1%
Students		84%	6%	1%	4%
1) Should all courses be evaluated by a mandatory policy of the Faculty Senate or Regents?					
			N	Yes	No
All Faculty Respondents			327	56%	37%
Assistant Professors			91	70%	30%
Associate Professors			101	60%	40%
Full Professors			135	50%	50%
All Student Respondents			848	86%	14%
Others			32	81%	19%
Undergraduates (Freshman to Seniors)			623	86%	14%
Masters Candidates in Sciences or Arts			153	86%	14%
Ph.D. Candidates			40	90%	10%
2) If a mandatory policy were in place should summaries of the results be published?					
All Faculty Respondents			321	56%	37%
Assistant Professors			93	66%	34%
Associate Professors			98	53%	47%
Full Professors			130	59%	41%
All Student Respondents			848	86%	14%
Others			32	81%	19%
Undergraduates (Freshman to Seniors)			623	86%	14%
Masters Candidates in Sciences or Arts			153	86%	14%
Ph.D. Candidates			40	90%	10%



QUESTIONNAIRE ON QUALITY TEACHING
TO FACULTY AND STUDENTS
AT THE UNIVERSITY OF NEW MEXICO

First, *please blacken in the appropriate demographic data.*

AGE	SEX	RANK IF FACULTY	LEVEL IF STUDENT	CAMPUS
51-UP ○ 41-50 ○ 31-40 ○ 25-30 ○ 18-24 ○	M ○ F ○	Full ○ Assoc ○ Assist ○ Instr. ○ Other ○	Ph.D. ○ MA or MS ○ Upper div. ○ Lower div. ○ Other ○	Main ○ Medical ○ Law ○ Other ○

Second, *should all the courses at the University of New Mexico be evaluated by students on a mandatory basis (say by a Faculty Senate and/or Regent's policy)?*

Yes ○ No ○

Third, *should some kind of summary of student evaluations of courses (if a mandatory policy were in place) be made available to ASUNM or some other body for publication?*

Yes ○ No ○

Fourth, *please rank each of the following elements (1-12), often associated with good teaching, first on Scale 1 for how important you think it is to good teaching and then on Scale 2 (four items) for how valuable you think each possible source of information on that element is.*

Scale 1: 4 = indispensable 3 = valuable 2 = useful 1 = weak 0 = useless
Scale 2: 4 = best source 3 = good source 2 = fair source 1 = weak source 0 = useless source

POSSIBLE ELEMENTS OF GOOD TEACHING	Use Scale 1: Importance	Use Scale 2: Undergraduates	Use Scale 2: Graduates	Use Scale 2: Peers	Use Scale 2: Self
1. Instructor's knowledge of subject.	1 ④ ③ ② ① ⓪	1 ④ ③ ② ① ⓪	1 ④ ③ ② ① ⓪	1 ④ ③ ② ① ⓪	1 ④ ③ ② ① ⓪
2. Research and publishing in subject area.	2 ④ ③ ② ① ⓪	2 ④ ③ ② ① ⓪	2 ④ ③ ② ① ⓪	2 ④ ③ ② ① ⓪	2 ④ ③ ② ① ⓪
3. Personal integrity and sincerity.	3 ④ ③ ② ① ⓪	3 ④ ③ ② ① ⓪	3 ④ ③ ② ① ⓪	3 ④ ③ ② ① ⓪	3 ④ ③ ② ① ⓪
4. Enthusiasm for subject and teaching.	4 ④ ③ ② ① ⓪	4 ④ ③ ② ① ⓪	4 ④ ③ ② ① ⓪	4 ④ ③ ② ① ⓪	4 ④ ③ ② ① ⓪
5. Effective use of illustrations, stories, etc.	5 ④ ③ ② ① ⓪	5 ④ ③ ② ① ⓪	5 ④ ③ ② ① ⓪	5 ④ ③ ② ① ⓪	5 ④ ③ ② ① ⓪
6. Being a listener, sensitive, empathic, etc.	6 ④ ③ ② ① ⓪	6 ④ ③ ② ① ⓪	6 ④ ③ ② ① ⓪	6 ④ ③ ② ① ⓪	6 ④ ③ ② ① ⓪
7. Presenting a challenge, motivating students.	7 ④ ③ ② ① ⓪	7 ④ ③ ② ① ⓪	7 ④ ③ ② ① ⓪	7 ④ ③ ② ① ⓪	7 ④ ③ ② ① ⓪
8. Encouraging appropriate interaction in class.	8 ④ ③ ② ① ⓪	8 ④ ③ ② ① ⓪	8 ④ ③ ② ① ⓪	8 ④ ③ ② ① ⓪	8 ④ ③ ② ① ⓪
9. Relevance of lectures, assignments, readings, etc.	9 ④ ③ ② ① ⓪	9 ④ ③ ② ① ⓪	9 ④ ③ ② ① ⓪	9 ④ ③ ② ① ⓪	9 ④ ③ ② ① ⓪
10. Encouraging appropriate interaction in class.	10 ④ ③ ② ① ⓪	10 ④ ③ ② ① ⓪	10 ④ ③ ② ① ⓪	10 ④ ③ ② ① ⓪	10 ④ ③ ② ① ⓪
11. Reasonableness of work instructor requires.	11 ④ ③ ② ① ⓪	11 ④ ③ ② ① ⓪	11 ④ ③ ② ① ⓪	11 ④ ③ ② ① ⓪	11 ④ ③ ② ① ⓪
12. Fairness in grading, examinations, etc.	12 ④ ③ ② ① ⓪	12 ④ ③ ② ① ⓪	12 ④ ③ ② ① ⓪	12 ④ ③ ② ① ⓪	12 ④ ③ ② ① ⓪

Figure 3. Questionnaire distributed at the University of New Mexico in academic year 1991-1992.

Although there are other factors besides a lapse of time and increasing experience with adult-level courses of instruction that may come into play in accounting for the trends toward greater agreement over time, as Figure 4 shows, in their progress from the Assistant Professor rank to Full Professor, faculty become less keen on the idea of being evaluated by students whereas students in progressing upward to the doctoral level seem to move in exactly the opposite direction, though less dramatically than the faculty. From the Assistant to the Full Professor rank there is a shift of 20 percentage points from 70% of the Assistant Professors favoring mandatory evaluation of courses whereas in progressing to the Full Professor level, the faculty seemed to lose faith in student evaluations dropping to only 50% favoring mandatory student evaluation of all courses. Among the students, those who progressed as typically part-time "non-traditional" students upward through full-time status to the doctoral level, the student view of mandatory course evaluation became more positive by 9 percentage points (moving up from 81% in favor to 90% in favor). Of course, it should be noted that the starting point for the student shift (at 81%) was nearer the ceiling on the scale at 100% than for faculty who started at a 70% approval rate and had more room to move toward the floor of the same scale.

In addition to the questions about mandatory evaluation and published summaries, student and faculty respondents were also asked to rank each of 12 "Possible Elements of Good Teaching" as 4 "indispensable", 3 "valuable", 2 "useful", 1 "weak", or 0 "useless", and to rate possible sources of information

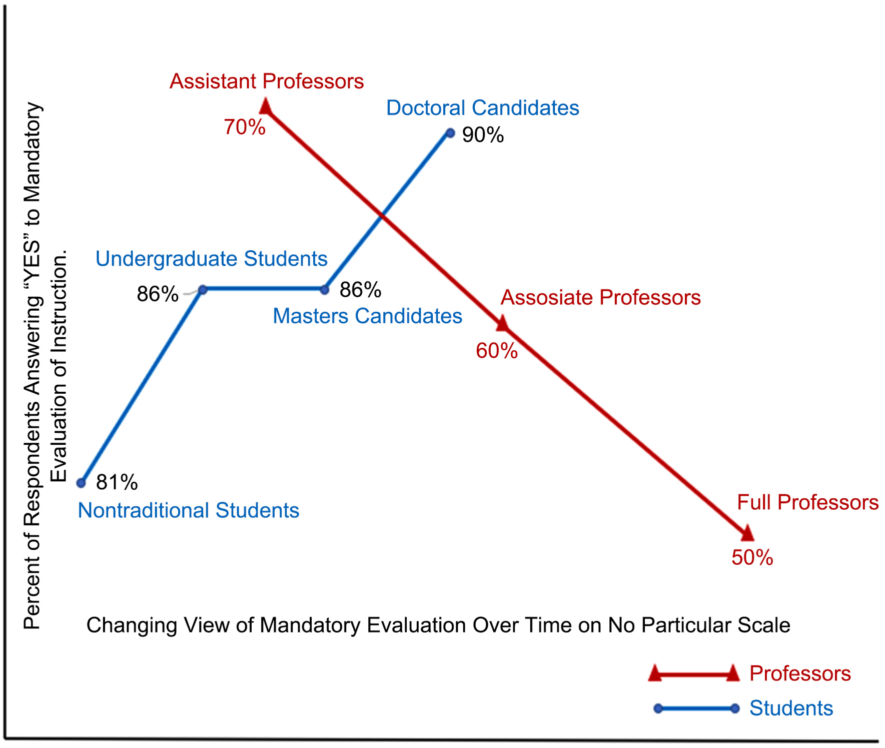


Figure 4. Percent of faculty (*N* = 327) and student respondents (*N* = 848) answering "yes" to a policy of mandatory course evaluation to be put in place by the Faculty Senate or Board of Regents of the University of New Mexico.

about courses of instruction (undergraduates, graduates, peers, and self) as 4 the "best source", 3 a "good source", 2 a "fair source", 1 a "weak source", or 0 a "useless source".

Table 2 shows the means and standard deviations as the various elements were ranked by students and faculty. Whereas Laupper and colleagues used a pair-wise approach to deleting missing cases, we used a list-wise deletion of missing cases so that the *N*s representing students (833) and faculty (304) are the same for all 12 of the "possible elements of good teaching". It is interesting to note that the very top ranked element—*ranked first by both students and faculty*—was knowledge of the subject-matter. It seems that almost everyone with a little experience in adult-level study or teaching of any kind intuitively appreciates the fact that the rest of the elements of good teaching are useless without an instructor having knowledge of the subject-matter[2].

[2] Oddly, the form used for student evaluation of instruction at the University of Louisiana during the 24 years of the author's service at that institution (mandatory for faculty but optional for students) has never included a scale referring to the instructor's knowledge of the subject-matter. While this important omission was pointed out to the responsible authorities, it was not corrected and has recently been accentuated by the new current practice of asking students to complete course evaluations before the latter third of the course is completed. Typically, course evaluations are done before final examinations have been taken and before the students' course grades have been reported to them. The policy in place is something like asking editors to review a paper for possible publication without regard for the competence of the author(s) and to make their decision (to publish or not) when about a third of the paper still remains to be written.

Table 2. Means and standard deviations for student (*N* = 833) and faculty (*N* = 304) rankings of the possible elements of good teaching.

Possible Elements of Good Teaching	Students		Faculty		Students	Faculty
	Mean	SD	Mean	SD	Rank	Rank
1) Instructors knowledge of subject.	3.767	0.481	3.842	0.383	1	1
2) Research and publishing in subject area.	2.575	0.901	2.704	0.799	12	12
3) Personal integrity and sincerity.	3.415	0.726	3.566	0.604	6	3
4) Enthusiasm for subject and teaching.	3.615	0.614	3.638	0.527	3	2
5) Effective use of illustrations, stories, etc.	3.019	0.819	2.803	0.754	10	10
6) Being a listener, sensitive, empathic, etc.	3.122	0.840	2.941	0.877	9	9
7) Presenting a challenge, motivating students, etc.	3.295	0.727	3.414	0.649	7	5
8) Organization, clarity, good syllabus, etc.	3.436	0.743	3.322	0.656	5	6
9) Relevance of lectures, assignments, readings, etc.	3.507	0.694	3.253	0.702	4	7
10) Encouraging appropriate interaction in class.	2.971	0.836	3.026	0.800	11	8
11) Reasonableness of work instructor requires.	3.228	0.777	2.793	0.844	8	11
12) Fairness in grading, examinations, etc.	3.627	0.646	3.526	0.669	2	4

Enthusiasm for the subject and for teaching that subject-matter was ranked second by faculty and third by students. Personal integrity and sincerity ranked third by faculty and sixth by students; fairness in grading, examinations, etc. ranked fourth by faculty and second by students, and so on. But all of the remaining elments would count for little or nothing if the instructor did not know the subject-matter. The nonparametric Spearman's rho as well as Pearson product-moment correlation for the last two columns in Table 2 is 0.83916. It is plain to see that the students and faculty at the mid-sized university sampled in this study agreed substantially, though not perfectly, on which elements of good teaching are most important. Nevertheless, it was supposed on the basis of sound theory that it should be possible by principal factoring to reduce the 12 scales to about three or four.

Whereas Laupper, Balzer, and Berger (2020) used confirmatory factoring to assess the relative agreement or disagreement between online and offline modes of evaluating the quality of instruction in Swiss vocational schools, at the University of New Mexico we used the simplest of exploratory factoring methods to reduce each of the two 12 × 12 correlation matrices to a minimum number of uncorrelated factors. With reliability estimates on the diagonal, each correlation matrix was factored and orthogonally rotated to give maximum loadings on the rows for factors with an eigenvalue at 1 or more.

The matrix for the 833 student respondents as shown in Table 3 yielded two orthogonally rotated factors of which the first is defined best, it seems, by whether the instructor encourages interactions (item 10, loading at 0.647), takes student input seriously (item 6, 0.606), uses good stories and illustrations (item 5, 0.575), is reasonable (item 11, 0.541), organized (item 8, 0.521), knows the subject matter (item 1, 0.506), and so forth. For students, the second factor was defined mainly by fairness (item 12, 0.693), relevance (item 9, 0.587), workload (item 11, 0.541), organization (item 8, 0.521), and knowledge of the subject (item 1, 0.506). Students are, apparently, concerned about how the instructor treats them and whether the instrution itself is meaningful to them.

The second factor analysis, addressing the results of responses from 304 faculty respondents, is shown at the right hand side of Table 3. Beginning with reliabilities on the diagonal of the initial 12 × 12 correlation matrix, as was also done with students, the factoring with a varimax orthogonal rotation yielded four factors with eigenvalues equal to or greater than 1. Factor 1 for faculty appeared to be about fairness (item 12, loading at 0.621), reasonableness (item 11, 0.556), and integrity (item 3, 0.542); factor 2 seems to be about challenging and motivating students (item 7, 0.554), encouraging interaction (item 10, 0.527), being a good listener (item 6, 0.510), and being enthusiastic (item 4, 0.474). Factor 3 seemed to be about course organization (item 8, 0.655) and relevance of the lectues and materials used (item 9, 0.501). Factor 4, got its only substantive loadings from knowledge of the subject (item 1, 0.466), and from research and publishing in the subject area (item 2, 0.403).

Table 3. Principal factors (in two separate analyses) for students and faculty on their respective rankings of the 12 possible elements of good teaching.

Possible Elements of Good Teaching	Students		Faculty			
	Factor 1	Factor 2	Factor 1	Factor 2	Factor 3	Factor 4
1) Instructor's knowledge of subject.	0.174	0.151	0.037	−0.040	0.151	0.466
2) Research and publishing in subject area. .aus	0.315	0.100	0.007	0.037	−0.081	0.403
3) Personal integrity and sincerity.	0.442	0.304	0.154	0.153	0.096	−0.038
4) Enthusiasm for subject and teaching.	0.490	0.369	0.078	0.474	0.110	−0.085
5) Effective use of illustrations, stories, etc.	0.575	0.197	0.243	0.406	0.297	−0.037
6) Being a listener, sensitive, empathic, etc.	0.600	0.268	0.479	0.510	0.099	−0.024
7) Presenting a challenge, motivating students, etc.	0.512	0.345	0.075	0.554	0.136	0.138
8) Organization, clarity, good syllabus, etc.	0.324	0.521	0.162	0.206	0.655	0.003
9) Relevance of lectures, assignments, readings, etc.	0.275	0.587	0.359	0.247	0.501	0.109
10) Encouraging appropriate interaction in class.	0.647	0.198	0.417	0.527	0.086	0.025
11) Reasonableness of work instructor requires.	0.412	0.154	0.556	0.306	0.160	0.105
12) Fairness in grading, examinations, etc.	0.171	0.693	0.621	0.033	0.183	0.060

Thinking through the factors revealed in student responses, it seems that students are most concerned with how the instructor treats them as interlocutors (Factor 1—items 10, 6, and 5), and secondarily with how the instructor delivers the subject-matter to them (Factor 2—items 12, 9, 8, 1). Faculty respondents, by contrast, seem to be a little more nuanced (as we should expect on account of the fact that they have also been students and still are in some sense): fairness, reasonableness of work required, integrity, and empathy (Factor 1—items 12, 11, 3, 6) form one component; challenging students, encouraging them, and having empathy for them accounts for the second component (Factor 2—items 7, 10, 6); being organized and relevant in lectures and assignments yields a third component (Factor 3—items 8, 9); and finally, the all important component of knowing what they are talking about stands out all by itself (Factor 4—items 1, 2).

It comes out that knowledge of the subject-matter (items 1 and 2) is at the foundation for both students and faculty, but, as might be expected, faculty seemed to be more concerned than their students about published research (item 2). Next to that, it seems that the way in which communication takes place, the articulation of methods from lectures to tests and grading (items 12, 11) as well as organization and relevance (items 8, 9)—all of these were regarded as a single factor for students (their Factor 2) but as two distinct components for faculty (Factor 2 for items 12, 11 and Factor 3 for items 8, 9). Also, whereas knowledge of subject-matter and published research emerges as a separate factor for faculty (Factor 4, items 1, 2), the existence of published research (item 2) seems to be largely ignored by the 833 students, most of whom were undergraduates, and knowledge of the subject matter (item 1) falls together with the factor that seems to be about how the course is presented to the students, e.g., with fairness (item 12), relevance (item 8), a reasonable workload (item 11), and evidence that the instructor knows the subject-matter (item 1).

Our last set of questions pertained to who might be best qualified to do the course evaluation for each of the "possible elements of good teaching". Depending on the aspect to be evaluated, it might be expected that some persons would be better qualified than others. With that in mind Table 4 shows that both faculty and student respondents, on the whole, see graduate students as the best evaluators on 5 of the 12 named elements. Faculty respondents (N = 304) regarded themselves as the best judges of their own knowledge of subject matter (item 1) and ranked their peers just higher than themselves as judges of their research and publications in the subject matter area (item 2). Students in general see themselves as the best judges of all the named elements except for research and publishing where they rated faculty peers above themselves. The faculty respondents, by contrast, rated themselves just barely lower than graduate students as judges of the relevance of their lectures, etc., whereas students, on the whole were quite consistent in rating themselves as the best judges of virtually everything about instruction except for faculty research and publications. Overall there was substantial agreement about the usefulness of the different sources

Table 4. Evaluating distinct sources of information about each of the possible elements of good teaching (faculty $N = 304$*, student* $N = 833$).

Element to Be Evaluated	Possible Evaluators										
	Undergrads (U)		Grads (G)		Peers (P)		Self (S)		Best Source		Best
	Student	Faculty	Student	Faculty	Student	Faculty	Student	Faculty	Student	Faculty	
1) Instructor's knowledge of subject area.	2.851	2.352	3.117	3.059	3.009	3.237	2.730	3.285	G	S	S
2) Research and publishing in subject area.	1.832	1.237	2.678	2.506	2.711	3.265	2.578	3.312	P	S	S
3) Personal integrity and sincerity.	2.939	2.858	2.981	3.067	2.974	3.087	2.900	3.063	G	P	P
4) Enthusiasm for subject and teaching.	3.266	3.368	3.103	3.308	2.695	2.577	2.984	3.174	U	U	U
5) Effective use of illustrations, stories, etc.	2.923	2.972	2.800	2.901	2.338	2.198	2.597	2.664	U	U	U
6) Being a listener, sensitive, empathic, etc.	3.079	3.032	2.942	3.075	2.625	2.320	2.692	2.609	U	G	U
7) Presenting a challenge, motivating students, etc.	3.007	3.051	3.026	3.308	2.492	2.399	2.727	2.822	G	G	G
8) Organization, clarity, good syllabus,	3.131	3.043	3.098	3.107	2.559	2.597	2.837	2.957	U	G	U
9) Relevance of lectures, assignments, readings, etc.	3.014	2.632	3.131	3.024	2.517	2.652	2.900	3.016	G	G or S	G
10) Encouraging appropriate interaction in class.	2.876	2.929	2.874	3.198	2.228	2.277	2.660	2.791	U or G	G	G
11) Reasonableness of work instructor requires.	2.956	2.285	3.014	2.664	2.492	2.462	2.622	2.818	G	S	G
12) Fairness in grading, examinations	3.135	2.870	3.156	3.083	2.643	2.597	2.783	2.984	G or U	G	G
Row means	**2.917**	**2.719**	**2.993**	**3.025**	**2.607**	**2.639**	**2.751**	**2.958**	**G or S**	**G or U**	**G**

of information concerning the quality of instruction on the various criteria presented. A noteworthy exception is the fact that faculty respondents viewed undergraduate students in the weak range as sources of judgments about faculty research and publications in the subject-area of instruction.

4. Determining the Course of Study: Its Subject-Matter, Methods, and Assessment Procedures

At this point it is important to say exactly what is meant by the subject-matter, methods, and assessment procedures in any course of study/training/coaching throughout this entire discussion here in PART ONE and also in Part Two[3]. Such a course consists of a sequence of events—e_1, e_2, e_3, ⋯, e_n—in the more or

[3] Although entertainment has some value in instruction, it is rarely the primary objective. It may supplement but is not normally the subject-matter of instruction. In late night talk shows, and stand-up comedy, juggling, dance, popular musical performances, and the like the entertainment may well be the main objective, but hardly in ordinary classrooms, remote asynchronous courses, college degree programs, or in training seminars for industry or business contexts. In most instructional contexts, at least in the ones of interest here, the goal is successful communication about some more or less well-determined (represented) subject-matter expressed and demonstrated through relevant performances of the teacher and often of students as well. In the cases of interest here, the main objective is communication of the subject-matter in an intelligible and consumable way that can be acquired and demonstrated by students along the way and also at the end of the course.

less shared experience of persons who complete the course. Based on the results of Laupper *et al.* [26], and proofs to be presented below here, it seems to matter little whether courses are completed synchronously, asynchronously, or in some mixture. According to the general consensus, as spelled out in detail in the "Quality Matters Rubric" and summed up in questionnaires for the evaluation of such courses (e.g., as in the many questionnaires constructed at the QuestionPro website [25], for instance), there must be some meaningful agreement (alignment) between the three major components in the sequence of events: 1) the subject-matter consisting of a series of successive experiences conversations, readings, written exchanges, demonstrations, performances, exercises, recitals, simulations, tests, etc., that form the storyline of the "course" of study, training, coaching, or whatever the orchestrated series consists of; 2) the methods of engagement by students experiencing and participating in the series—readings they work through, lectures they attend, paraphrases, summaries or critiques they produce, performances they model or improve, demonstrations they experience or produce, skills they practice, drills they perform, tests they take, essays they write, speeches presented or critiqued, recitals presented or critiqued, etc.; and 3) measures attained as expressed in scores on tests or graded performances consisting of binary, or n-ary multiple choice items, open-ended fill-in-the-blank items, written or oral essays, or other performances assessed and ranked for quality by classmates, instructors, and/or subject-matter experts, instructors, coaches, or some combination of the foregoing.

5. Any Independent Series versus Any Real Dependent Series

Noting that the classrooms, the computers, the internet itself, and all the people in the world with all their real and imagined fears, problems, etc., are already real—the reader may be wondering how the mathematical proofs referred to in the abstract of this paper can be applied to minimize failures and maximize successes in the planning, execution, and evaluation of university/college curricula/degree programs or business/computerized training courses for certification, licensure, and the like? It is this last very general question that I am addressing in both parts of this paper.

6. Defining "Independent" Events Abstractly

Interestingly, the applications to be made here of the Kolmogorov proofs about probability theory also afford a straightforward definition of any series of "independent" events along lines that Kolmogorov himself would later develop to an advanced level in what is known today as "Kolmogorov complexity" [27] [28] [29]. His chain rule of complexity involves a strict computational approach to obtain the shortest program to reproduce a string of possible symbols in a given code, and to distinguish between meaningful and random sequences in a general and useful way. The most interesting applications, as I see them, are the ones that have been developed in theories of biochemical resonance in genetic, epigenetic, proteomic, and related bioinformatic phenomena [30]. However, in the present application to instruction and to its evaluation, we can keep our thinking very simple. In the measurement of the efficacy of a course of study, or in the evaluation of instruction/training in general, we can contrast the measured impact on intelligent human beings of a course of study with that of an unthinking robot (or some number of them) making random decisions about test items in a course of study not informed by meaning at all. Let the robot (or a group of them) generate random "independent" choices in response to meaningful questions with the sole caveat being that the robot knows nothing of the relevant meanings—nothing of the subject-matter. For the purposes of this paper I am interested in applications that reason from effects backward to causes along the lines of the Bayesian inference, and from causes to expected effects in the ordinary experimental contexts of instructional courses of study.

7. The Definitions of Bayes 1763 and Kolmogorov's Axioms 1933

In 1763, a paper by Thomas Bayes [31], augmented and explained after his death by his friend Richard Price, appeared in the *Philosophical Transactions of the Royal Society*. There Bayes laid down seven definitions which could be read as rules of procedure in reasoning about the likelihood of the occurrence of certain events. Here I want to apply that reasoning to the measurement of instructional success, but with a significant amplification of the Bayesian system with respect to what he refers to as "failed" events—ones not observed to happen in some spatio-temporal context or other. Such failures, as I will show, are completely uninformative, except in instances where the known absence of one event guarantees the positive presence of another.

To show the dependence of everything Bayes had to say, and of mathematical proofs in general on the theory of true narrative representations (TNR-theory; [10] [11]), we only need to examine closely the fourth definition of Bayes in his list of seven: "an event is said to be *determined* [my italics] when it has either happened or failed". Bearing in mind that the absence of any event in experience, or our failure to notice such an event, proves nothing unless the missing event is supplanted by a present event that is part of an independent series in the Kolmogorov sense, the indefeasible proof that the *determination* of the occurrence of *any actualized event whatsoever depends on a TNR in every possible instance follows from the question*, how can we know if any given event has occurred? To make such a *determination* about actual facts, in every conceivable instance requires, as I have already proved [10] [11], and as I will summarize below, a TNR. In the quoted matter that follows I will only modernize the ancient spellings of Bayes and I will add a few words in square brackets for clarification of what I understand his meaning to be:

> Problem. *Given* the number of times in which an unknown event [one as

> yet *undetermined* with respect to the present time of the observer] has happened and failed [not happened]: *Required* the chance that the probability of its happening in a single trial lies somewhere between any two degrees of probability that can be named [here he does not mean merely to "name" those degrees, but to fix them on the basis of prior knowledge of how frequently the event in question has happened in the past].

Section I.

Definition: 1) Several events are *inconsistent* [incompatible or cannot occur together at the same time], when if one of them happens, none of the rest can.

2) Two events are *contrary* [exclusionary opposites] when one, or [the] other of them must [happen]; and both together cannot happen.

3) An event is said to *fail* [during some time segment of observation] when it cannot happen; or, which comes to the same thing, when its contrary [exclusionary opposite, like heads coming up instead of tails on a coin toss; or any of 35 other possibilities when double sixes appear on a throw of a pair of dice] has [already] happened.

4) An event is said to be determined [as a known fact; oddly not italicized by Bayes!] when it has either happened or failed [a *determination* that absolutely depends on an act by a competent observer who represents the happening of the event or the failure to happen of that event via a true narrative representation, a TNR, that can only be constructed, as I will prove below, after the fact].

5) The *probability of any event* is the ratio between the value [the gain if it happens as contrasted with the possible loss if it fails to happen] at which the expectation depending on the happening of the event ought to be computed, and the value of the thing expected upon it's [*sic*, its] happening [the bet won or lost, or the contract fulfilled or not, or the benefit obtained or cost avoided; this "value" can only be estimated by some competent observer (or a group of them) capable of *determining* the outcome of the expected event via a TNR, whether it happens or fails to happen].

6) By *chance* I mean the same as probability. [Here a peculiar equivalence is established. Why not say "chance" everywhere when the concept is needed? And presumably, we can substitute "likelihood" for either one in the theory of Bayes. In any case, the definition is clear enough: by "chance" Bayes understands some estimation of the likelihood of an event.]

7) Events are independent when the happening of any one of them does neither increase nor abate the probability of the rest. [Conversely, we may infer that dependent events are such that the happening of one increases or decreases the likelihood of the happening of the other.]

7.1. Previewing TNR-Theory

Next, it is crucial to emphasize that to make sense of any of the "Definitions" laid down by Bayes, the positive "*determination*" of the happening of any single event, or of any series of events that might inform all the rest of his reasoning, is absolutely essential. That is to say, unless we are competent to identify occasions when a certain event or series of events actually occurs (or has occurred in the past), we are powerless to make any judgments whatever about the likelihood—"chance" or "probability" of such events relative to the past, present, or future. What Bayes does not make clear is that for what he calls "failed" events, ones not observed to happen—unless some other independent event is known to take the place of the heads that does not appear on the face of the coin, or the double sixes that do not appear on the faces of the dice—a failure to observe a particular event at a given time and place, *tells us nothing*. Such "failed" events are completely uninformative to a limit of containing absolutely no information whatever.

To prove: a *determination* of the occurrence of any event requires a TNR. The competent observer cannot merely imagine the occurrence of any event of interest, but must perceive it faithfully (truly) when it happens in the real world of space, time, and matter. However, the record of any such experienced event is of no use to anyone but the person who experienced it unless it is noted and reported in some intelligible way comprehensible to other intelligent persons. Because any event that is so *determined* can only be reported after it has occurred, the required record absolutely must be of a *narrative* kind—as in a true story told after the fact. Further, in order for the report to have the power to inform other intelligent persons, so they can reconstruct the nature of the event (even if they did not perceive it), the report must enable the construction of a faithful image of the event. To be faithful to the event in the required sense, the narrative must be true—it must not report anything that did not occur and whatever it does report must have occurred as reported. Therefore, to *determine* any event that actually occurs and is observed to occur by an intelligent observer, an intelligible and true re-presentation of the narrative kind, a TNR must be produced—QED, which was to be demonstrated.

Moreover, the existence of any such TNR as is required to *determine* the occurrence of any event whatsoever absolutely requires the faculties of one or many intelligent observers capable of discerning the difference between the occurrence of any event (success) as contrasted with the occurrence of other events. Moreover, such TNRs can only serve the purposes of Bayes, or any scientist, observer, teacher, trainer, student or what-have-you, to the extent that those TNRs are fully intelligible to similarly gifted persons. At the outset, it must be noted that if success is *determined* in a TNR representing the event that has actually occurred, it produces the sort of agreement between the TNR and the persons who understand it, that can only be described properly as unitary. If we must put a number to it, it has to be the number one. *For example, in the tossing of a coin, if it should come up heads and this is reported in a TNR, any competent observer who understands the TNR has sufficient shared information about the fact(s) referred to so as to be able to reconstruct it (or them) as completely as required in order to* determine *which of the two possible events has succeeded* *and which has failed. That is to say, the TNR stands, relative to the observers who understand it and to the fact*(*s*) *it* determines, *in relatively perfect, unitary agreement*. The informativeness of any TNR, therefore, must be judged as = 1.

By contrast, failure to observe an expected event may come about for many reasons and here are just a few of them in no particular order and without any limit on the list: 1) *the event may have actually occurred but the time frame for observation may have been too brief*; 2) *the event may have occurred but not have been observed because the observer was not paying attention or did not have a powerful enough telescope or microscope to see the event, or because it had become too dark, too noisy, etc. to observe the event*; 3) *the observer may have blinked or been sufficiently distracted at just the moment when the event occurred so it was not noticed*; 4) *the event may have occurred just before or just after the time frame allowed so that it occurred but was not observed*; 5) *the observer may have fallen asleep during the time frame when the event occurred*; 6) *the event may have been observed and recorded but the record may have been lost in transit and never recovered*; 7) *the event may have occurred in a place not under observation*; 8) *the pre-requisites for the occurrence may have been missing, no coin to toss, so no heads or tails to appear*; 9) *some part of the pre-requisites may have been missing, the coin might have been blank on one or both sides*; … *and so on and so forth with no end in view. So, how can we put a reasonable estimate on the informativeness of an event that does not occur*? *There is a reasonable estimate of the informativeness of such a non-occurrence, and it must be zero. The informativeness of a failure to observe an expected event tells us exactly nothing*. The information in a failed experiment = 0.

7.2. Connecting All the Foregoing with Instruction

The teacher and student alike are hoping for and trying to achieve success: so a correct answer to a test question, or success in demonstrating a skill, always tells us more than failure. Our concern is mainly about "success" in the Bayesian sense. To cut to the chase and say how all the reasoning to be brought to bear pertains to ordinary courses of study or instruction, success in answering a test question, leaping over a hurdle, landing an aircraft safely, completing a course of study, and that sort of thing, can only be expressed or understood in a TNR that agrees as perfectly as it purports to agree with the event or event sequence it purports to represent. As for failures, we want to avoid them and to do so, the bottom-line is that we absolutely require TNRs. If a student succeeds in producing the correct answer, or demonstrating the desired skill, failure is excluded, and cumulative successes as they occur with increasing regularity, can with growing confidence as successes accumulate, be attributed to KSAs acquired. Failures, by contrast, can have many causes, but can, over the long haul with successful instruction, be expected to converge toward zero. As a result all of our reasoning comes down to determining how success (in the Bayesian sense as applied to tests and measurements of instruction) can be maximized while the likelihood of its contrary (failure) is minimized.

7.3. Anticipating PART TWO of This Paper

Part Two of this paper further develops the mathematical reasoning from Part One and shows how it relates to actual measures of instructional success in radically diverse test formats and in repeated presentations of the same subject-matter by the same instructor to comparable samples of students drawn from the same population. Haertel's 2013 theory of "value-added" by instruction [32]—directly measured by improved student test scores, all else being held equal—is applied. He argued that if the students' abilities and motivation to learn, the subject-matter difficulty and the tests used to measure it, and the instructor(s) doing the presentations, could all be held approximately equal, instructional success should be directly measurable by the proximity to mastery (100% shared information) of all the students progressing along the way, and especially at the end of the course. Generalizing to the repeated presentations of any course where the instructor aims to improve the "alignment/agreement" of the three critical course components (subject-matter, methods of presentation, and performance testing) from one presentation to the next, value-added (or lost), all else being held equal, should appear as a progressive increase (or decrease) in student scores on comparable measures relative to how close they come across time to the limit of complete mastery (100% shared information).

8. Conclusions from PART ONE

Mathematical proofs pertaining to the central limit theorem and probability by Kolmogorov, amplifying proofs by Pólya and by Bayes, show that successes in communication over time—particularly those deliberately engineered in a course of instruction where teachers and designers aim to produce a sequence of true narrative representations—must (all else being held equal) cumulatively progress toward a limit of unity of shared information. That is, as the interlocutors accumulate information through determinate TNRs of the subject-matter—as found within the bounds of a series of episodes in a course of study, training exercises, workshops, or whatever—the increasingly shared information must trend in the direction of the limit of 100% of the information in the course. That theoretical limit can in fact be viewed as a pragmatic definition of the "mastery" of that course of study.

Bearing all that in mind, Part One of this paper confirms that there is pervasive agreement (convergence toward unity) about even the most abstract qualities of instructional successes and how to achieve them. In Part Two we consider the most concrete elements of subject-matter measured by radically different forced-choice test items as contrasted with open-ended short answer test items, and we also perform a series of experimental measures of value-added by better aligning instructional components and methods in actual empirical contexts. The data reveal that reliability and validity of radically different kinds of test items are enhanced by better communication of the subject-matter, and that progress toward mastery of targeted KSAs is, unsurprisingly, but dramatically, also improved by better alignment/agreement of the key elements of the course of study.

Acknowledgements

Most importantly, I thank Carolyn Kennedy, Teresa M. Meehan, and Kathy McCoy for help with student surveys at UNM. Also, thanks to Ellen Laupper, and colleagues for sharing material from their study of online (new school) versus offline (old-school paper and pencil) evaluations of instruction by students. They generously provided an English translation of their questionnaire along with unpublished tabled data that was most helpful. I am grateful for their generosity and appreciate the quality of their work. Daniel Broudy, James Lyons-Weiler, Brenden D. Oller, Christopher A. Shaw, and anonymous readers at OALIB helped me purge errors and infelicities. Whatever of those may remain, are my own.

Conflicts of Interest

The author declares no conflicts of interest.

References

[1] Dewey, J. (1929) Nature, Communication, and Learning. In: Hayden, D.E. and Alworth, E.P., Eds., *Classics in Semantics*, Philosophical Library, New York, 265-296.

[2] Badon, L.C., Oller, S.D., Yan, R. and Oller, J.W. (2005) Gating Walls and Bridging Gaps: Validity in Language Teaching, Learning, and Assessment. *Studies in Applied Linguistics and TESOL*, **5**, 1-15.

[3] (2021) Pinky and the Brain [Internet]. Wikipedia.

[4] Riemann, B. and Clifford, T.W.K. (1873) On the Hypotheses Which Lie at the Bases of Geometry. *Nature*, **8**, 14-17, 36-37. https://doi.org/10.1038/008036a0

[5] Weinberger, E.D. (2002) A Theory of Pragmatic Information and Its Application to the Quasi-Species Model of Biological Evolution. *Bio Systems*, **66**, 105-119. https://doi.org/10.1016/S0303-2647(02)00038-2

[6] Gernert, D. (2006) Pragmatic Information: Historical Development and General Overview. *Mind and Matter*, **4**, 141-167.

[7] Oller, J.W. (2013) Pragmatic Information. In: Marks II, R.J., Sanford, J.C., Dembski, W.A. and Gordon, B.L., Eds., *Biological Information*: *New Perspectives*, World Scientific Publishing Company, Singapore, 64-86. https://doi.org/10.1142/9789814508728_0003

[8] Maleeh, R. (2014) Pragmatic Information as a Unifying Biological Concept. *Information*, **5**, 451-478. https://doi.org/10.3390/info5030451

[9] Zhao, M., Liu, T. and Chen, F. (2018) Automatic Processing of Pragmatic Information in the Human Brain: A Mismatch Negativity Study. *Neuroreport*, **29**, 631-636. https://doi.org/10.1097/WNR.0000000000001009

[10] Oller, J.W. (2010) The Antithesis of Entropy: Biosemiotic Communication from Genetics to Human Language with Special Emphasis on the Immune Systems. *En-*

tropy, **12**, 631-705. https://doi.org/10.3390/e12040631

[11] Oller, J.W. (2014) Biosemiotic Entropy: Concluding the Series. *Entropy*, **16**, 4060-4087. https://doi.org/10.3390/e16074060

[12] (2020) Knowledge, Skills, and Abilities [Internet]. Wikipedia.

[13] (2021) Quality Matters [Internet].

[14] Eisner, E.W. (1999) The Uses and Limits of Performance Assessment. *Phi Delta Kappan*, **80**, 658-660.

[15] Kim, H. (2004) Task-Based Performance Assessment for Teachers: Key Issues to Consider. *Studies in Applied Linguistics and TESOL*, **4**, 1-5.

[16] Ritter, J. (1996) Transcript of Crash Shows Controller Error/Review Reveals Poor English "Over and Over". *USA Today*, 7A.

[17] Day, B. (2002) Proposed ICAO Proficiency Requirements in Common English. Part I: The Need for English Language Proficiency Requirements and the Role of the International Civil Aviation Organization. 8*th International Aviation English Association Seminar*, Warsaw, September 2002, 44.

[18] Tajima, A. (2004) Fatal Miscommunication: English in Aviation Safety. *World Englishes*, **23**, 451-470. https://doi.org/10.1111/j.0883-2919.2004.00368.x

[19] International Civil Aviation Organization (2004) Language Proficiency (Presented by the Secretariat). In: *Fourteenth Meeting of the APANPIRG ATM/AIS/SAR Sub-Group*, International Civil Aviation Organization, Bangkok, 5.

[20] Yan, R. (2009) Assessing English Language Proficiency in International Aviation: Issues of Reliability, Validity, and Aviation Safety. Lambert Academic Publishing, Koln.

[21] Yan, R. (2013) Assessing the English Language Proficiency of International Aviation Staff. In: *The Companion to Language Assessment*, American Cancer Society, New York, 484-496. https://doi.org/10.1002/9781118411360.wbcla050

[22] Eberhard, D.M., Simons, G.F. and Fennig, C.D. (2019) Ethnologue: Languages of the World [Internet]. 22nd Edition, SIL International, Dallas.

[23] Broudy, D. (In Press) Transgressing the Logic of the New Tyrannical Normal: An Essay on the Struggle for Reason and Freedom. *International Journal of Environmental Research and Public Health*.

[24] Deng, R. and Benckendorff, P. (2021) What Are the Key Themes Associated with the Positive Learning Experience in MOOCs? An Empirical Investigation of Learners' Ratings and Reviews. *International Journal of Educational Technology in Higher Education*, **18**, 9. https://doi.org/10.1186/s41239-021-00244-3

[25] (2021) 12 Amazing Course Evaluation Survey Questions [Internet]. QuestionPro.

[26] Laupper, E., Balzer, L. and Berger, J.-L. (2020) Online vs. Offline Course Evaluation Revisited: Testing the Invariance of a Course Evaluation Questionnaire Using a Multigroup Confirmatory Factor Analysis Framework. *Educational Assessment Evaluation and Accountability*, **32**, 481-498. https://doi.org/10.1007/s11092-020-09336-6

[27] Gnedenko, B.V. and Kolmogorov, A.N. (1968) Limit Distributions for Sums of Independent Random Variables [Internet]. Addison-Wesley, Boston.

[28] Ryabko, B.Ya. (1986) Noiseless Coding of Combinatorial Sources, Hausdorff Dimension, and Kolmogorov Complexity. *Problemy Peredachi Informatsii* (*Problems of Information Transmission*), **22**, 170-179.

[29] Ryabko, B., Astola, J. and Gammerman, A. (2006) Application of Kolmogorov Complexity and Universal Codes to Identity Testing and Nonparametric Testing of

Serial Independence for Time Series. *Theoretical Computer Science*, **359**, 440-448. https://doi.org/10.1016/j.tcs.2006.06.004

[30] Petoukhov, S.V. (2018) Structural Connections between Long Genetic and Literary Texts [v2]. https://doi.org/10.20944/preprints201812.0142.v1

[31] Bayes, T. and Price, R. (1763) Essay towards Solving a Problem in the Doctrine of Chances. *Philosophical Transactions of the Royal Society of London*, **53**, 370-418.

[32] Haertel, E.H. (2013) Reliability and Validity of Inferences about Teachers Based on Student Test Scores. Center for Research on Human Capital and Education, Research and Development, Educational Testing Service, Princeton, The National Press Club, Washington DC.
https://www.ets.org/Media/Research/pdf/PICANG14.pdf