Open Access Library Journal
2021, Volume 8, e7729
ISSN Online: 2333-9721
ISSN Print: 2333-9705

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

John W. Oller

Institute for Pure and Applied Knowledge, IPAK-EDU.org, Pittsburgh, Pennsylavania, 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 TWO). *Open Access Library Journal*, **8**: e7729.
https://doi.org/10.4236/oalib.1107729

Received: July 3, 2021
Accepted: July 25, 2021
Published: July 28, 2021

Abstract

PART ONE incorporated probability theory into the theory of true narrative representations (TNR-theory). The proofs show that interactional successes in a course of study must trend toward 100% shared information. Failures converge on zero. PART TWO compares robotic responses to forced-choice test items (an independent series of events) against university students whose responses are dependent on what they learned. The proofs are also applied to experimental measures of what Haertel in 2013 referred to as "value-added" by instruction. When all factors except the course presentation can be held constant, gains in test scores can be taken as measures of value-added by instruction. Such gains as predicted by the proofs are confirmed (1) for radically diverse methods of testing the same subject-matter across time, and (2) for seven iterations of a course, tested by nearly identical test items aimed at the same subject-matter, but with improved alignment/agreement across the critical components of subject-matter, methods of presentation, and procedures of assessment.

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

For precept must be upon precept, precept upon precept; line upon line,

line upon line; here a little, and there a little… (Isaiah 28:10, King James Version of the Bible)

1. Generalizing Kolmogorov's Proofs to Real Instruction

To define what he meant by "independent" (random) events from the abstract mathematical point of view, Kolmogorov [1] created a list of axiomatic constraints grounded in set theory but taking into account the reasoning of Bayes 1763 [2], and Pólya 1921 [3] [4] in his central limit theorem. The axioms of Kolmogorov enabled his elegant proofs about probability theory in 1933. Especially interesting were pages 69 - 70 in [1] about why certain combined probabilities must converge either to 1 or 0.

To develop his arguments, Kolmogorov required an event series such that if any one of the events would actually occur, all the others of that series would be excluded. To illustrate, consider the tossing of a coin. It cannot come up heads and tails at the same time, but it must come up one or the other. Or, take the throwing of a single die: the face of it can only show 1, 2, 3, 4, 5, or 6, never more than one of these outcomes at the same time. Similarly, a pair of dice can show exactly 36 possible outcomes—each of them "independent" of the others because only one such pair of numbers can come up on any given throw. Even if the coin were weighted or the dice loaded, Kolmogorov's type of independence would remain.

Of course, in the real world, the tossing of a coin, throwing of a pair of dice, or the dropping of a feather and a bowling ball from a height—none of these, is an event that happens by accident. It is done by someone most commonly to decide whether or not to take some risk, how many spaces to move in a game of monopoly, whether the feather and the bowling ball will fall at the same rate (as predicted by Galileo and Einstein), or if they may fall at different rates (as predicted by Aristotle[1]), and so on and so forth. Even the independent series of events dealt with by Kolmogorov, following Bayes and Pólya, are always embedded to the extent that they are actually *determined* in experience (by deliberate actual experiments). To be represented they require TNRs as proved in PART ONE. Next, let us look more closely at how such *determinations* are made.

2. "Likelihood" and Real-Life Events

In the experiment of tossing a coin, for instance, there are exactly 2 possibilities on each repetition of the experiment. In the dropping of the feather and the bowling ball there are exactly three possibilities: 1) they can fall at the same rate (as Galileo expected and as Einstein explained later on in greater detail), or 2) the bowling ball can fall faster (as Aristotle seems to have incorrectly predicted), or 3) the feather can fall faster (though *no one ever predicted that counter-intuitive possibility*).

[1] Though to give him his due, could he have taken account of the resistance of the air in which case a bowling ball would fall quite a lot faster than the feather? But, alas, in a vacuum as Einstein predicted correctly, they fall at the same rate.

2.1. A Series of *Independent* Events

For the tossing of a coin, the probabilities assigned to the two possible outcomes (events), in the abstract are heads, 1 chance in 2 possible outcomes, and tails 1 chance in 2, or ½ + ½ = 2/2 = 1, or a 100% probability of getting either heads or tails on every toss and a 0% chance of getting neither. If we ask what are the odds of getting say heads (*h*) on the first toss and tails (*t*) on the second (or pick any of the 4 possible outcomes, *hh*, *tt*, *ht*, *th* that you like), the odds are 1 against 3 in favor and 3 to 1 against whichever outcome we might predict. For the chance of getting 10 heads in a row, we must multiply ½ times ½ 10 times to obtain the odds in favor of the pre-determined sequence at 1 chance in 2^{10} possible sequences. This is the same as saying that there is only 1 of 1024 possible independent sequences of 10 coin tosses that can result in a win for someone who bets on that particular sequence as predicted in advance. The odds against that bet are 1023 possible outcomes against the 1 bet on. The same sort of reasoning can be applied to a sequence of events such as, for example, throwing double sixes three times in a row. The odds in favor the person placing such a bet would be 1 in 36^3 possible outcomes. The odds against the betting person would be exactly 746,495 to 1.

So, summing up, by mathematical induction, we may infer that for all possible series of independent events, the sum of probabilities of the possible independent events must equal unity, 1, and the possibility of getting a particular sequence of 2, 3, 4, … or *n* outcomes in that series converges on zero as the series is extended toward infinity. Also, the combined cumulative probability of randomly obtaining a particular sequence of events can be obtained by multiplication in the way just illustrated for the coin and the dice. The result hoped for, in a longer and longer pre-determined sequence becomes more and more unlikely by chance as the number of predicted outcomes increases.

2.2. A Series of *Dependent* Events

Translating the foregoing to ordinary testing in a course of study, if the test consists of some combination of *k* binary and *j n*-ary multiple-choice questions, the chance of stumbling onto some sequence of correct answers by chance is equal to the *k*th power of ½ to account for the binary items, multiplied by the *j*th power of 1/*n* to account for the non-binary items with *n* choices each. If short answer questions or essays, or performances and demonstrations to be judged somewhat in the manner of essays, are included in any of the sequence of test items for a course of study, the denominator immediately begins to approximate infinity so that raising it to the power of the number of such open-ended items presents a hurdle so high that no accidental effort in the real world of space and time will ever be sufficient to leap over it. A robot giving random answers to any sequence of items that inclued open-ended answers or essays will invariably fail. But what about a person capable of learning as the sequence of events in a course unfolds? Will such a person fail even if the tests are very difficult? Keep in mind

Riemann's proof (described in PART ONE) about a slight positive curvature as an analogue for the modicum of understanding every intelligent person has to start with. Riemann showed that no matter how big the universe might become, if space has a slight positive curvature, the space must remain finite. Likewise, a little understanding can account for the whole subject-matter if extended over enough lessons across a sufficient period of time.

3. The Real World as Problematic

Kolmogorov expressed the peculiar problem of "the concept of independence" this way:

> one of the most important problems in the philosophy of the natural sciences is—in addition to the well-known one regarding the essence of probability itself—to make precise the premises which would make it possible to regard any given real events as independent. [2, p. 9].

Eventually, he would appeal to randomized choices made by a computer in order to extend his theory to real-life events [5] [6]. With that in mind, from a biosemiotic perspective [7] [8] [9] [10] [11] it is possible to give a pragmatically satisfactory definition of a random choice in a series of independent possible events without stretching our imagination much at all.

3.1. Real Life Random Responses

Suppose we think of a student in a real course at a university, in a business training course, or an interlocutor in whatever situation in the real world who comes to a forced choice with no knowledge of what is being asked. I have occasionally had this kind of reaction to the sorts of trivia questions sports fans or Hollywood groupies thrive on about batting averages, players in the NBA, the names of actors in a sit-com I have never watched, and so forth. Suppose the individual knows only that the desired response is a binary choice (true or false, yes or no). Or it might be an n-ary choice. Or it could be an open-ended fill-in-the blank question where some unknown number of answers (many of them in some cases) would satisfy the questioner. In a still more difficult case, the request might call for an explanation of something the respondent knows nothing about.

In a worst case scenario, the respondent may not even know the language in which the question is posed. In such real-life scenarios, the difficulty is similar to the one faced for want of information, by the pilot who asks in Mandarin Chinese, while reading the English phrase from his digital screen, "What does 'PULL UP! PULL UP!' mean?" All this seconds before the jumbo jet he is flying crashes into a runway in Tokyo killing himself and 260 others on board [12] [13].

A student who has no background in the subject-matter and no basis for choosing between true and false choices in a series of items. Also, such a student has no basis to prefer any one of the multiple alternatives offered in a multiple-choice test item, and much less could such an individual pull an appropriate technical term or phrase, still less an essay explaining a process, experiment, theory, or whatever from the thin air. For a person who has not read the assigned materials, listened to the lectures, and so forth, or who does not know the language of instruction, the situation is something like that of a person who cannot read a note of music being asked to hum the melody written on a sheet of music, or to play an instrument never practiced. How likely is any such challenge to meet with success?

3.2. A Robotic Solution

I started to write here that "if Kolmogorov were still living", he might find the foregoing description of a possible series of real world "independent" events too ill-defined to qualify as sufficiently similar to the purely abstract series of axiomatically defined independent events in the proofs constructed by Bayes, Pólya, and himself. But then I discovered that a good while after he wrote his 1933 treatise on probability theory, he had posed a completely general form of exactly the same problem that I want to consider here with reference to several rather special cases of information communicated between different parties about some subject-matter defined as I have already suggested in the preceding Part One of this paper, Section 4 titled "Determining the Course Subject-Matter, Methods, and Assessment". In 1965, Kolmogorov [14] posed this problem:

> Actually, it is most fruitful to discuss the quantity of information "conveyed by an object" (x) "about an object" (y)… The real objects that we study are very (infinitely) complex, but the relationships between two separate objects diminish as the schemes used to describe them become simpler. While a map yields a considerable amount of information about a region of the earth's surface, the microstructure of the paper and the ink on the paper have no relation to the microstructure of the area shown on the map.

Similarly the sound sequences, characters, words, phrases, and sentences in a given language—think of the sequence of representations in a course of study as Kolmogorovo's "object x" conveying information about that subject-matter of that course which is Kolmogorov's "object y"—have no *determinate* relation to whatever that subject-matter may be except for intelligent persons who are able to understand the language of the course and map the sequence of representations "x" onto subject-matter "y". Kolmogorov wrote:

> In practice, we are most frequently interested in the quantity of information "*conveyed by an individual object x about an individual object y.*"

To this characterization if we added a third term, the person expressing the information, and a fourth, the person interpreting it after it is expressed, we arrive at the appropriate level of complexity for the question, what is "shared information"?

> …how much information is contained in *War and Peace* [the 1869 novel by Leo Tolstoy]. Is it reasonable to include this novel in the set of "possible novels," or even to postulate some probability distribution for this set? Or, on the other hand, must we assume that the individual scenes in this book form a random sequence with "stochastic relations" that damp out quite rapidly over a distance of several pages? Actually, we are just as much in the dark over the fashionable question of the "quantity of hereditary information" necessary, say, for the reproduction of particular form of *roach*.

3.3. Putting It All in Ordinary Language

Kolmogorov asks how difficult would it be to develop a "program *p* for passing from an object *x* to an object *y*". Putting the same problem in terms of TNR-theory, the question is whether it is possible given a TNR to pass from its surface-form *x* to a valid construction of its factual subject-matter *y*. The program *p* required to do this, or the reverse, to go from *y* to *x*, must consist of the sort of indexical relations found in valid interpretations of the TNR at issue—all of which require persons who know the language of that TNR. Cutting to the bottom-line we come to what was later termed "Kolmogorov complexity" [6] [14] [15].

3.4. Mutual or Shared Information

From such thinking the notion of "mutual" or "shared" information invariably comes up, and eventually, information itself is defined as the antithesis of entropy (see my discussion in 2010 [7]). The idea is succinctly described in a Wikipedia article [16] about "shared" or "mutual information":

> Intuitively, mutual information measures the information that X and Y share: It measures how much knowing one of these variables reduces uncertainty about the other. For example, if X and Y are independent, then knowing X does not give any information about Y and vice versa, so their mutual information is zero. At the other extreme, if X is a *deterministic* [my italics on this word and all derivatives of the verb determine] function of Y and Y is a *deterministic* function of X then all information conveyed by X is shared with Y: knowing X *determines* the value of Y and vice versa. As a result, in this case the mutual information is the same as the uncertainty contained in Y (or X) alone, namely the entropy of Y (or X). Moreover, this mutual information is the same as the entropy of X and as the entropy of Y… Mutual information therefore measures dependence in the following sense: I(X; Y) = 0 if and only if X and Y are independent random variables.

3.5. Making the Invidious Comparisons

Suppose we put a robot (or several of them), in the place of one or more students who are bluffing their way through some course of study without bothering to understand any of it at all. Suppose further that the robots are provided with no pragmatic knowledge of the subject-matter in the requisite course of study, and no knowledge of the language in which the questions are posed—except for the power to randomly arrange surface forms in that language into strings of syntactically well-formed sentences and to choose randomly between various alternatives presented in test questions. It makes no difference whether we have in mind a university course or an industry/business training seminar of some sort. Regardless what subject-matter content and format for presentation of the material may be chosen for the course, suppose only that at the end of each module of content, a series of test items are presented that require forced choices (binary or *n*-ary), or linguistic forms to fill in blanks with suitably short or longer essay-type sequences of symbols that the robot must draw at random from the surface-forms of the presented subject-matter but that intelligent students are asked to study, interpret, discuss, and to learn.

3.6. Meeting the Independence Requirement

Given that all the responses of the robots must be strictly random (from a base of zero information), the sequences of responses from one question to the next will always meet Kolmogorov's requirement of "independence" and the likelihood of "mutual information" being greater than zero trends rapidly to zero itself. By contrast, the likelihood of shared information between the instructor, the subject-matter, and the intelligent students who study the various representations of whatever the subject-matter material may be, is good to start with because they understand the language in which the subject-matter is represented. Also, the likelihood that their shared information will increase over time is assured as the course progresses, provided only that the students have some intelligence and stick with the drill, and that the instructor(s) add subject-matter information as the course progresses. The measure of mutual information must progress away from zero and toward a theoretical limit of unity as Kolmogorov's TNR-amplified theory requires. To the extent that the representations, interpretations, and tests of the subject-matter, involve TNRs, the agreement between intelligent persons must trend invariably toward unity.

4. Communication in the Real World

It is true, of course, that imaginary conversations and private thoughts take place in every intelligent individual, but even those imagined events require a certain amount of real time and all of them take place only with the assistance of a real body located in the real world. As we have already seen, classrooms and the contexts of all communication interactions are already in the real world. There is no need, to tear down any barriers between the components of real-life interactions, because there are no real barriers to tear down.

4.1. Three Independent Orders of Agreement Exist

To assure successful communication between interlocutors is merely necessary to optimize as much as possible the degree of agreement/alignment between 1) the representations of the defined subject-matter of instruction—the curriculum of that particular course of study—as presented through 2) a series of real interactions designated for learners as events, e_1, e_2, ⋯, e_n for any single learner (or any group of learners) consisting of encounters with the readings, lectures, recordings, demonstrations, exercises, activities, and so forth leading to 3) the tests, all of which should ultimately define and correspond as faithfully as possible to actual performances expected of persons who have completed some or all of the course.

4.1.1. First-Order Agreement, the Subject-Matter: What Are the Facts?

The first order of agreement is knowledge of the subject-matter—defined by readings, lectures, exercises, games, performances, or whatever the course may consist of. The prediction following from Kolmogorov's proofs and their amplifications is that such first-order-agreement must increase as interlocutors succeed in gaining experience with that subject-matter irrespective of whatever it may be and quite independently of whatever the interlocutors (students, teachers, coaches) may believe about the propositions, claims, etc., contained in the subject-matter.

Gains in the knowledge of the subject-matter may be greatly assisted along the way by good methods of communication as contrasted with not so good ones, and by better instruction as contrasted with not so good teaching, but the attained knowledge of the subject-matter, in the final analysis, is about facts that are quite independent of the methods of instruction or the attributes, including the beliefs, of the instructor and the students in the course. For example, whether or not Noam Chomsky claims to believe that fictional worlds are just as important to language learners as the common real world of experience, or that some commentator believes the US border with Mexico should be closed or open, or that global warming is a certainty or a hoax, or that aluminum is good for you if you eat or inject it—none of these facts involve whether the interlocutors in the course agree with such statements or believe the persons making them. First-order agreement does not require any commitment at all to the beliefs of other persons except about whatever the subject-matter of the course of instruction is.

4.1.2. Second-Order Agreement: What Must I Do? Interactions Required?

The second order of agreement involves whatever sequence of interactions are required of and experienced by students working through whatever the combination of methods of instruction for the course may be. That sequence of events may be thought of as a chronology of encounters with whatever methods the instructor(s) has (or have) provided and whichever ones learners may have engaged with first, second, and so forth right up to the end of the course of study. It matters not at all to the subject-matter itself whether the interactions accounting for the sequence of events—e_1, e_2, ⋯, e_n—take place in an old-school classroom, a prison setting, in a gymnasium or swimming pool. They may involve face-to-face and hand-to-hand contexts, or occur on the internet, or in some mix of synchronous and asynchronous encounters. Irrespective of the particular methods of delivery and interactions required (and irrespective of the proportion that are required or received), second-order agreement must increase over time to the extent that participants actually engage with the subject-matter through those methods in a finite sequence of whatever events constitute the course of study for that particular student or group of them. However, just as opinions about the subject-matter are at a level above the subject-matter, opinions about whatever methods of instruction have actually been experienced are also above the experience with those methods.

4.1.3. Third-Order Agreement: What Enables Success? Evaluating Instruction

Finally, third order agreement is the sort that is achieved as students, performers, athletes, coaches, instructors, and the like, gain experience over time with the sorts of series of interactions constituting multiple courses of study or training. Again, as with first-order and second-order agreement, third-order agreement (convergence toward a unity of shared information) is largely independent of any particular subject-matter and of any particular methods of presenting it. The question at this level is what are the requisite abstract properties of successful instructional interactions? What enables communication in general to work for me and my classmates if there are any attributes of instructors or qualities of instruction that work for us as contrasted with those that don't work? What qualities or competencies must the teacher/coach possess and employ to help me, and others like me to succeed? What interactions must be required of students/trainees in order to optimize our chances for success? Again, if the particulars of the subject-matter and the methods of interaction are set aside, the Kolmogorov proofs and their amplification according to pragmatic information and the theory of TNRs show that third-order agreement, like the first and second orders of agreement, must also trend toward a limit of unity as intelligent (and truthful) interlocutors (students and instructors) gain experience with courses of instruction over time.

5. Dependent Event Sequences Are Not Random

Because intentional communications are never random, in my initial studies of how children acquire their first language [17] [18] [19], it was not uncommon for theoreticians to insist that the input enabling the discovery of the meanings of words, phrases, sentences, and so on in a natural language is random [20] [21]. What made such a statement coming from Noam Avram Chomsky so eye-poppingly strange was that it came from one of the 10 most quoted intellectual of all time and the only one of them still living [22] [23] [24]. Nonetheless, it is not quite true. Chomsky took a wrong turn in making syntax the driver in his theories, as John Sowa ([25], p. 1037) noted in 2018. By neglecting the conceptual aspects of human understanding—the abstract semantic meanings and concrete pragmatic meanings that connect linguistic strings with the facts of real world experience [17] [19] [26] [27] [28]—Chomsky [20] [29] [30] could embrace the illusion that the linguistic strings presented to babies are random. More recently, while incorporating aspects of those domains [31], he continues to maintain [32] [33] that the real world holds no status above that of fictional worlds invented by the imagination. Yet, TNR-theory [7] [8] [34] and its precursors [35] [36] [37] [38] [39] proved the real world not only has special status, but that without it *no fictional worlds whatsoever can possibly be constructed* [40].

Peirce [37] produced a succinct version of the argument in his proposal to the Carnegie Institution:

> …every proposition whether it be believed, doubted, asserted, commanded, or put interrogatively, supposed, etc. essentially represents itself to represent an absolute reality [the real world]… This reality is not in any respect constituted by being represented… If a proposition represents that reality and represents it rightly in whatever respect it represents it, the proposition is true. If the proposition does not represent the absolute reality or in any respect represents it wrongly, it is false.

5.1. Kolmogorov Refutes Chomsky

Kolmogorov's complaint about how difficult it is to find an independent (random) series of events in the real world applies in spades against Chomsky's assertion that the input to a child learning any language is random. If the baby is hungry and cries, someone, usually the baby's mother, feeds the baby. If the diaper needs changing, it gets changed. Day follows night with regularity and routines are repeated along with meaningful representations of whatever is going on in producing those routines. They consist in the vast majority of instances of TNRs. The sequence of such events continues from the remembered past into the present experience, and it extends toward the future. As a result the continuous narrative stream enables a cumulative growth of knowledge about what has happened in the past, what is going on now in the present, what is about to happen in the future, and what happened even before the present series of events in the individual's experience began [40]. That is to say, the individual living through this series of TNRs is able to draw inferences not only about events that range beyond the present, e_p, designated as the series $e_1, e_2, \cdots, e_p$, but extending into the future of that present, $e_{p+1}, e_{p+2}, \cdots$, and even backwards to the past preceding the individual's lifetime designated as the amplified series… $e_{p-2}, e_{p-1}, e_1, e_2, \cdots, e_p, e_{p+1}, e_{p+2}, \cdots$

5.2. Information Involves Truth and Entropy Its Absence

As argued long ago by Peirce [35] [41] [42] and a little later on by Tarski [38] [39], meaningful sign systems universally depend on access to representations that happen to be "true" in the least burdened sense of that term. With respect to language acquisition, which if it were blocked in all cases would preclude the existence of any natural languages, I later proved in several different ways [7] [8] [34] that all language acquisition (and the decipherment of any meaningful sign system whatever) depends on the existence of TNRs. Late in his life, at the age of 60, Peirce summed up his own argument along these lines by describing as "a discovery of no little importance" expressed eventually in his logical proofs that "all knowledge without exception comes from observation" [37].

6. Integrating Kolmogorov's Proof into TNR-Theory

Without working through all the details the relevant constructions of TNR-theory can be summed up in a series of ranked inequalities flanked on both sides by approximate equations using Kolmogorov's 1 and 0 as follows:

$$1 \approx \text{TNRs} > \text{fictions} > \text{errors} > \text{lies} > \text{nonsense} > \text{erasures} \approx 0$$

Summing up the significance of the whole ranked series of systems, at the extreme left, TNRs are the gold standard. They alone are relatively perfect among all those possible representational sequences that have any meaning (any information about anything). The subject-matter of every TNR delivers all that is claimed by its surface-forms (think of them as Kolmogorov's x) and the linking of one to the other through the bi-directional indexes (Kolmogorov's program p) that connect the surface-forms with the subject-matter (Kolmogorov's y) in every TNR such that there are no erroneous links. This is the same as saying that the subject-matter can be re-constructed from the symbols alone, and vice versa. Putting the relationship in terms of Kolmogorov's third quantitative definition of information [14], actually *mutual information*, which comes near its ideal form in any TNR where the information "conveyed by an individual object x about an individual object y" is so nearly complete that object y can be constructed through program p if we are only given object x.

6.1. A Genetic Example

The first genetic example of that sort of successful Kolmogorov-type reconstruction may have been the one referred to by Plothe in 2010 ([43], p. 181). Allegedly it involved the complete re-construction of the complex hormone oxytocin from its genetic description. In theory, TNRs enable such reconstructions because of the high degree of mutual information shared by their requisite components. The sequence of symbols must agree as perfectly as they purport to agree with the indices that connect them to their factual subject-matter. Similarly, the subject-matter 1) must deliver through the indexes 2) all that is claimed by the sequence of symbols 3). In fact, it has been proved that the three components of every TNR form a relatively perfect trinity such that the information contained in any one of the parts is contained in all three and in each of the other two components (see http://www.johnoller.com/tnr-proofs/). Notably, however, TNRs are the only sign systems among all that are possible that have these peculiar perfections [7] [8] [34].

6.2. Fictions Are Less Complete than TNRs

Fictions are constructed by derivation from TNRs by allowing some part of the subject-matter of in some TNR (or complex of them) to be replaced by an imagined projection of that subject-matter. That process of imagining invariably involves the logical operation of abstraction. It is that same process that enables the construction of hypotheses that can be manipulated mathematically or empirically. In fact, the process of generalization, critical to reasoning about non-finite sets was explicitly noted by Kolmogorov. He wrote: "Infinite fields of probability occur only as idealized models of real random processes" ([1], p. 15). The fiction, then, is less complete than the TNR from which it may be derived because the derived fiction contains some imaginary part that any real subject-matter with which it may be associated cannot deliver. However, as Kolmogorov noted "if reasoning which utilizes the probabilities of … ideal events leads us to a *determination* [my italics] of the probability of an actual event [in the ideally generalized set]…, then, from an empirical point of view also, this *determination* [my italics] will automatically fail to be contradictory" ([1], p. 18).

It is interesting that Kolmogorov does not offer any proof that the resultant "*determination*" cannot contradict any truth, but TNR-theory shows why. All TNRs, no matter how they may be discovered, must agree with all the rest of them. They cannot contain any contradictions, or else, they would not be true. At the same time, TNRs can comprehend all other types of representations, whereas no other kind, only TNRs, can adequately represent any other TNRs as well as any fictions, errors, lies, nonsensical strings, and all the precursors up to complete erasures.

6.3. Errors, Lies, Nonsense, and Erasures Are Increasingly Less Perfect

With derived errors there is a further corruption beyond the imaginary projection required to produce a fiction from a TNR: in the instance of an error, some fictional part of the error must be mistaken for a TNR. To correct the error it would be necessary to replace that fictional part with an actual TNR. With lies the corruption takes a further step away from ordinary truth: the lie must involve a known error dressed up to deceive interpreters into thinking it is not a mere fiction, nor yet an error, but rather a TNR. In nonsense strings in any given symbol system, the surface-form remains to some discernible extent and yet the indexical connections with any possible subject-matter are to the extent of the nonsensicalness corrupted to a greater or lesser degree. The proof that nonsense strings in general are less intelligible than lies is that to qualify as such, lies must maintain intelligibility—that is, unlike nonsense, lies must retain a meaningful resemblance to TNRs allowing at least a partial projection onto a factual state of affairs. However, if that projection were to contain no contradictions of the facts that the lie purports to represent, it would not be a lie, but rather a TNR. Finally, it is possible to erase all vestiges of the starting point, say, in a TNR devolved to some fiction, error, lie, or nonsense, in which case the resulting blank space, to the extent of the destructive erasure, has approximately zero information as suggested by the approximate equation at the right hand side of the formula given above at the top of this section.

6.4. Determinacy, Connectedness, and Generalizability

Among the unique perfections (logical completenesses) of TNRs are the traits of *determinacy*, *connectedness*, and *generalizability* only found in TNRs [7] [8] [34]. The upshot of all the reasoning about the nature of ordinary truth is that without TNRs to enable the discovery of the meanings of signs and sign systems, there can be no meaningful signs at all and therefore no meaningful sign systems. Language acquisition, and the deciphering of any existing meaningful sign system, depends exclusively on TNRs. In natural languages and in genetics, epigenetics, proteomics, and so forth, the existence and deciphering of such systems likewise depends utterly on the existence of TNRs in those sign systems.

7. Meaningful Abstractions Are Anchored in the Real World

With reference to *n*-ply extended manifolds, Riemann [3] described the necessity of finding a starting place in the real world in this way:

> I have in the first place… set myself the task of constructing the notion of a multiply extended magnitude out of general notions of magnitude. It will follow from this that… space is only a particular case of a triply extended magnitude. But… the properties which distinguish space from other conceivable triply extended magnitudes *are only to be deduced from experience* [my italics; again, an anticipation of TNR-theory and its proofs]. (pp. 1-2)

Because the dimensions and shape of any solid object, or the theoretical space it occupies in however many dimensions we may wish to take into account can be varied without any necessary constraints other than those of our imagination, it is plain to see that fictional (imaginary) manipulations of TNRs, are the *sine qua non* of the invention of the fictions constituting mathematical constructions in general and, therefore, of mathematical reasoning in its entirety. Whereas it is possible, and in mathematical reasoning, even necessary to abstract away from the constraints of the particular real measures and the factual context of any given object, or any complex of object/event relations in the real world, for the purposes of generalization, all such constructions require some starting point, or some complex of starting points, from which we can make sense of our abstractions and generalizations.

7.1. Abstracting from Something Like a Diagram

For such reasons as Riemann expressed [3]—ones worked out in amazing detail by Peirce [27] [41] [42]. Peirce insisted that all mathematical reasoning begins with something like a diagram (also see his existential graphs as explained by Sowa [44] [45])—a picture or sketch of an actual object or complex of relations between objects that can form the basis for the generalization to countless similars. At the basis of any such diagram there must be one or more TNRs. An extraordinary example, from Einstein [46] is the projection of shadows onto a plane from a disc contained on the surface of a sphere as drawn in **Figure 1**. The great physicist created his diagram elaborating on what Riemann had previously written arguing that even the slightest positive curvature of space would ensure that the universe is approximately spherical and finite in its expanse. Riemann's statement was translated and published in 1873:

> …if we assume independence of bodies from position, and therefore ascribe to space constant curvature, it must necessarily be finite provided this curvature has ever so small a positive value. [3] (p. 10)

In drawing **Figure 1**, Einstein argued that a shadow cast upon a flat 2-dimensional surface from a disc, one of a finite number covering the whole surface of the sphere, can be made to grow to infinite size by increasing the distance of the surface E from the sphere K showing that the sphere is not infinite in its extent and yet is unbounded because the projected shadows can be indefinitely expanded: In his concluding remarks Einstein wrote somewhat enigmatically:

> In this way, by using as a crutch the practice in thinking and visualization which Euclidean geometry gives us, we have acquired a mental picture of spherical geometry. [46]

7.2. *Determination* Requires a Real World

To obtain any of the requisite ideas for Kolmogorov's probability theory and its extensions—no matter how abstract they may be made out to be—the determination of just one such event in the slightest degree, as well as its connection with similar events in a finite series consisting of n such events, and the enablement of the abstract comprehension of Kolmogorov's generalization to a hypothetically infinite series of such events, or to a perfected would not even be interpretable were it not possible to relate it to "the actual world of experiments" [1]. Interestingly, Reimann [40] implied this point and it was taken seriously by none other than Albert Einstein [46] who wrote:

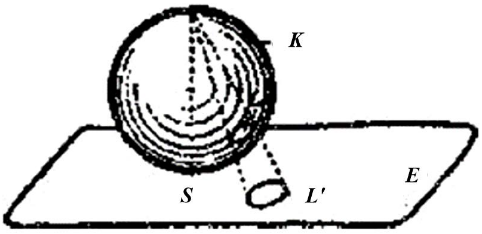

Figure 1. A drawing from Einstein used in his Lecture on "Geometry and Experience" at a meeting of the Prussian Academy of Sciences in 1921, illustrating his argument for an unbounded and yet finite surface on a sphere showing its limited finite capacity and yet its unbounded nature at the same time. (Einstein here used an argument first constructed more succinctly by Riemann in 1873.)

> …the investigator in another department of science would not need to envy the mathematician if the propositions of mathematics referred to objects of our mere imagination, and not to objects of reality. (p. 1)

7.3. Independent Events Do Not Occur in Ordinary Instruction

Whereas Kolmogorov said that making "precise the premises which would make it possible to regard any given real events as independent" was "beyond the scope" of the book he was writing—it is not difficult now, with the benefit of the additional proofs concerning the nature of ordinary truth from Peirce [35] [36] [41] [42], Tarski [38] [39], Kolmogorov himself and his followers [6] [14] [15] [47] [48], and more recently TNR-theory [7] [8]—to draw a comparison between the non-independence of a series of TNRs in ordinary instruction with the sort of randomized series that would be generated by a robot with nothing more than a superficial knowledge of the need for choices ranging from 1) a binary true/false-type question, or 2) any *n*-ary multiple choice question, or 3) any open-ended question where the robot could construct a random sequence of surface forms up to some pre-determined number of characters, using no more than the syntax Chomsky had in mind, to form a syntactically acceptable sequence to fill the blank, or 4) to construct a longer essay of some arbitrary length supposedly to address a question that the robot does not understand.

8. A Finite Series of TNRs in Tests about a Given Subject-Matter

Keeping our discussion strictly within the limits of Kolmogorov's proofs about finite sets, suppose we compare robotic performances against intelligent adult students in, for instance, a series of 14 courses completed by 492 intelligent adults with an average of 350 items, of which 35 were binary and 315 were quintenary items, and where on the average adult students never answered fewer than 0.75 nor more than 0.94 of the items correctly. Actually, in each of those courses additional items were included requiring short fill-in-the-blank answers or essays of up to 250 words, but no robot with zero information in the subject-matter can produce any intelligible answers to such open-ended questions, so let us concentrate our attention strictly on multiple-choice questions with between 2 and 5 alternatives.

8.1. Testing Robots with Forced-Choice Items

The problem for one or more robots producing purely random choices in response to test items, trying to accidentally stumble to the right sequence of TNRs, is insurmountable. Being as generous as possible to the robot(s), suppose we use the lower end of the average performance of the weakest class where 0.75 of the 350 items were answered over the semester-long course of study. Because the score achieved by any given student, or robot, is indifferent to just which items are answered correctly, we may take any 0.75 of the items we like for the desired invidious comparison. Given that 0.75 of 15 binary items yields about 11 of those and 0.75 times 315 items of the quintenary type rounds up to 236 of those, the robot operating with exactly zero information must choose the correct sequence of binary choices 11 times straight and the correct sequence of quintenary choices 236 times to get within a half point of the desired 0.75 mark. For any single robot, the odds against the correct sequence of answers for just 11 binary choices would be $1/2^{11}$ or 1 against 2050. But the robot would also be obliged to answer 236 quintenary items where the odds against would be 1 to 5^{236} which comes to 9.0557 times 10^{164}. To combine the 11 binary items with 236 quintenary items the odds against the robot getting the correct sequence for that many items would be $(1/2^{11})*(1/5^{236})$ or 1.8564 times 10^{168}.

If we put a million robots to work on the test in question, we would only reduce the odds against any one of them matching the performance of an average student in the lowest performing group to 1 against 1.8564 times 10^{162}. The upshot is that a million robots answering multiple-choice items randomly would be hopelessly outmatched by just any average student. Independent sequences of random events as produced by uninformed robots must tend to converge on zero—no shared information at all.

8.2. Experimental Tests of Value-Added Gains

There are many ways to test the convergence to unity in successful communications as predicted by the extension of Kolomogorov's proofs. In this section and the next, two experimental studies are presented. The first experiment examines two quite different formats for testing knowledge of the same subject-matter (multiple-choice versus open-ended test items), and the second approach critically examines 7 iterations of the same course of study where the goal across those iterations was to adjust the presentation from semester to semester in a way that would improve what the Quality Matters Rubric describes as "alignment"—the agreement between the subject-matter, methods of presenting it, and the methods of assessing student uptake through measurable performances.

8.2.1. Integrating Testing and Teaching

In both of the empirical studies to be examined the test items were cross-referenced to the assigned reading material in the assigned eBook in the same chronological order as the relevant information was presented in that text. In the first study to be discussed below, the aim was to examine the degree of variance overlap between the multiple-choice (MC) questions drawn from a pool of 687 items with concomitant open-ended (OE) questions taken from a pool of 351 of the latter items aimed at the same subject-matter. Students were given all the questions and all the answers in advance although in constructing OE tests the wording could be changed so as to sometimes require a different correct answer other than the one already provided. The essential feature of every question was its being linked to the place in the text where the facts necessary to *determine* an answer to that particular question were found. It was supposed that a large item bank—one consisting of 1038 items (687 MC items plus 351 OE items)—would make it difficult (if not impossible) to memorize all the questions without coming to a fairly deep understanding of the underlying concepts, relations, theory, research, etc., in the assigned subject-matter of the course. Nevertheless, there were a couple of critics who argued that the test questions should be kept secret right up until the time students were to be challenged with them. Personally, I argued in favor of making all the tests in all the courses completely transparent. Given that known dichotomy, the following philosophical argument from Solomon may help to clarify my application of the amplified Kolmogorov proofs to the instructional issues at stake here.

8.2.2. Connecting the Kolmogorov Proofs to Solomonic Wisdom

In the ancient book of Proverbs, Solomon recorded (and possibly invented) the following proposition which is here presented in the majestic language of the *King James Authorized Version* of the Bible:

> There is that scattereth, and yet increaseth; and there is that withholdeth more than is meet, but it tendeth to poverty. (Proverbs 11:24, *KJV*)

The communicator/teacher who takes Solomon's advice and "casts their bread upon the waters" (Ecclesiastes 11:1)—that is, who makes the subject-matter and all the ways of looking at it as accessible as possible—will tend to see increasing dividends as knowledge grows over time and communication effectiveness increases along with it. Learning pays compounded interest. It is not a "zero-sum" game. James Lyons-Weiler captures the essence of "scattering" in the Solomonic sense at the Institute of Pure and Applied Knowledge which aims to bring "knowledge and people together in the pursuit of successes in research" (not failures, by the way, which are notoriously uninformative).

8.2.3. Comparing MC and OE Test Pairs across Time in the Same Course

With all of the foregoing in mind, in the summer of 2019, the author sought and obtained approval from the University of Louisiana Institutional Review Board for assessing some of the predicted convergences in two consecutive but separate presentations of his course on human anatomy and physiology. The objective was to assess the validity of multiple-choice (MC) versus open-ended (OE) formats applied systematically in multiple pairs of MC and OE tests in the same subject-matter over the course of two semesters. The correlations between respective pairs of MC and OE tests covering the same subject-matter would usually be understood as indices not merely of reliability but of test validity on account of the fact that the methods of testing are radically different.

8.2.4. Examples of MC and OE Test Items Addressing the Same Subject-Matter

Here are two pairs of sample items to illustrate the fact that the same content, more or less, can as easily be addressed in the MC format as in an OE format:

MC example 1: The term *anatomy* is a compound from the Greek morphemes *ana-* (ανα-) and *-temno* (-τέμνω) which combined in ανατομία mean to ____.

1) restore,
2) cut up,
3) heal quickly,
4) examine closely,
5) prevent disease.

OE example 1: The term *anatomy* is a compound from the Greek morphemes *ana-* (ανα-) and *-temno* (-τέμνω) which combined in ανατομία mean to____.

MC example 2: By showing how Pouchet's experiments were contaminated by germ-laden particles of dust, Pasteur refuted Pouchet's claim to have proved___.

1) the non-existence of microbes without parents,
2) the spontaneous generation of microbes,
3) the interdependence of trillions of microbes,
4) all of the above,
5) none of the above.

OE example 2: By showing how Pouchet's experiments were contaminated by germ-laden particles of dust, Pasteur refuted Pouchet's claim to have proved___.

The main difference between MC and OE items, in any case, is the fact that the former only require a choice between alternatives *provided by the test-writer* while the latter require *the person taking the test to supply a correct answer* by producing an appropriate string of symbols that expresses the requisite facts as discussed in reading materials and other resources. Ideally, however, if these contrasting formats of test construction are validly addressing the subject-matter that has been expressed in some manner or other in the course materials, lectures, discussions, tutorials, animations, and so forth, there should be a significant and substantial correlation between pairs of tests consisting of MC and OE items cumulatively covering the successive parts of a course. Moreover, correlations should be stronger for more comprehensive tests as knowledge of the subject-matter increases over the course of study and the learning begins to pay compounded interest on account of the human capacity to learn to learn (so-to-speak).

8.2.5. Participants

All the students in the experimental study at hand ($N = 77$), were undergraduate majors in speech-language pathology and audiology. In compliance with a curriculum more or less mandated for accreditation of university programs by the American Speech-Language-Hearing Association across the nation, all such students must take an introduction to the science of human anatomy and physiology. Two distinct groups of undergraduates were taught and tested over the requisite subject-matter at a mid-sized, fully accredited American university (namely, the University of Louisiana at Lafayette): the first group consisted of 36 majors who took the introductory course in the fall of 2018, and the second group was comprised by 41 students who took the course in the spring of 2019. There were 5 students, not included in the study, who completed the course but, owing to absences, did not complete all of the MC or OE test pairs. Only students who completed all of those pairs of tests were included. In that way, the design conforms to the requirements of a high-powered repeated-measures approach ([49], pp. 516-542) where individual performances are orthogonalized by the fact that no individual or group of individuals is ever compared against a different individual or a different group of individuals.

8.2.6. The Context in which the Tests Were Administered

There were 28 face-to-face class meetings of 75 minutes each followed by a 29th of 150 minutes for the final examination. The first pair of tests, MC1 and OE1, comprehensively covered material preceding class meeting 7 when that pair was administered. That pair of tests consisted of 50 items in the MC format and 30 items in the OE format. A second pair, MC2 and OE2, with 100 and 30 items, respectively, but including about twice as much content and drawn from all the items covered up to the mid-term, was administered on meeting number 14. Then, at the end of the course, a third pair of tests, MC3 and OE3, with 200 and 30 items, respectively, including questions drawn from the full data bank of 1038 items covering the whole course, was administered as the final examination. In addition, a singleton cumulative OE test consisting of 42 items arranged in three distinct parts covering (A) the articulators and resonating cavities of speech production, (B) key brain landmarks, and (C) cranial nerve pairs and their functions was administered at meeting 21.

8.2.7. Testing One General and Three Specific Hypotheses

From the mathematical proofs cited above, many empirically testable hypotheses can be derived. Among them is the general hypothesis that better communication of the content to be taught (all else being held equal) should result in better reliability, validity, and interpretability of test results obtained. That general hypothesis should hold for both the MC and OE items illustrated above. In addition, there are three specific hypothesis to be tested:

> Hypothesis 1: Provided successful communication is occurring throughout the class meetings, the correlations between successive pairs of MC and OE tests should increase from occasion 1 to 2, and from 2 to 3. (The corresponding null hypothesis is that correlations between successive MC and OE pairs will not increase across occasions.)

Hypothesis 2: Given the increase in coverage and test length obtained by adding MC1 to MC2, and OE1 to OE2, the correlations between composite scores created by adding the corresponding MC and OE tests from occasion 1 to 2, and those of 2 to 3, as well as the corresponding MC and OE tests of all three occasions 1, 2, and 3 (*i.e.*, MC1 + MC2, and OE1 + OE2, and so forth) should exceed the correlation between any single pair of MC and OE tests. (The corresponding null hypothesis is that correlations between successive MC and OE cumulative tests will not exceed those for one or several single pairs.)

Hypothesis 3: The measured agreement between MC and OE tests should be greatest in the correlation between the sum of all the MC tests with the sum of all the OE tests. (The corresponding null hypothesis is that the correlation between the sum of all MC and the sum of all OE tests will not exceed the other correlations computed between individual pairs of MC and OE tests or the correlations between the composites with fewer cumulative items.)

8.2.8. Results of the Analyses

Table 1 gives aggregated statistics for the 36 participants from the spring semester of 2018 combined with the 41 participants in the fall of 2019 ($N = 77$). Reported are the means, standard deviations, and estimated lower bound reliabilities for all the tests administered. The crucial pairs of multiple-choice (MC) and open-ended (OE) tests used to challenge experimentally the three hypotheses presented above, were administered at meetings 7, 14, and 29. The table, however, also includes descriptive statistics for an independent open-ended test administered at meeting 21 over the human articulatory anatomy, brain landmarks, and cranial nerve pairs. That test covered (A) the articulators and resonating cavities, (B) brain landmarks highlighted in the course, and (C) cranial pairs of nerves involved in innervating speech and related processes. However, this ABC test, of which students were apprised often before the 21st meeting, when it was administered, elicited an average performance near the ceiling on that test. The mean expressed as a proportion of 100 points was 92.93 suggesting the test was easy. However, it was without doubt the most demanding of all the OE tests included in Table 1. Students took it very seriously, studied the material thoroughly, and the majority of them mastered it sufficiently well to make the variance on this test at $s^2_{\text{ABC}} = 9.38^2 = 88.016$ almost 37 times smaller than the geometric mean of the variance on the other three 30-item OE tests at $\left(s^2_{\text{OE1}} + s^2_{\text{OE2}} + s^2_{\text{OE3}}\right)/3 = 3251.260$.

Whereas the estimated lower bound of reliability for the ABC test at 0.536 is the least of the reliability estimates in Table 1, contrary to appearances, because of the reduction in variance owed to the diligence of students in appreciating the importance of the ABC subject-matter to their curriculum, those scores come closer than any others to the limit of agreement at 100%. As they near that limit, of course, the reliability falls off toward a misleading zero. Assuming only that the highest goal of instruction is mastery, the convergence toward unity on the ABC test must be judged as 37 times better than that for any of the other OE tests, all of which were nonetheless reliable and valid in their own right.

In Table 2, we find the required correlations and the coefficients of determination for testing the three hypotheses stated above. Figure 2 shows the predicted near monotonic increase in the agreement, measured by correlations, between scores on the first MC and OE pair of tests administered at meeting 7, with r = 0.506; the second administered at mid-term (meeting 14), r = 0.519; and the final examination at meeting 29, with r = 0.540.

Table 1. Aggregated statistics ($N = 77$) for MC and OE tests completed by students in fall 2018 ($n = 36$) and spring 2019 ($n = 41$) in a course on human anatomy and physiology.

Test and Test Type	Test Pairs in Order, Meeting Number When Test Was Administered	Mean per 100 Points Possible	Standard Deviation	Lower Bound of Estimated Reliability*
Multiple Choice Test 1 (MC1)	Pair 1, 7	86.86.	13.59	0.959
Open Ended Test 1 (OE1)	Pair 1, 7	72.06	20.13	0.711
Multiple Choice Test 2 (MC2)	Pair 2, 14	85.42	9.98	0.910
Open Ended Test 2 (OE2)	Pair 2, 14	57.12	20.22	0.721
Articulatory Anatomy, Brain Landmarks, and Cranial Nerve Pairs (OE$_{ABC}$)	Unpaired OE$_{ABC}$, 21	92.93	9.38	0.536
Multiple Choice Test 3 (MC3)	Pair 3, 29	75.86	12.98	0.961
Open Ended Test 3 (OE3)	Pair 3, 29	74.01	23.97	0.879

*For the MC tests the lower bound is the computed Kuder-Richardson 25 for the largest sub-sample (n > 18 < 28) of participants tested at a particular administration in the spring semester 2018 ($n = 36$) and in the fall semester of 2019 ($n = 41$). For the paired OE tests, the estimate is the square root of the correlation with the concomitant MC test, and for the unpaired OE test (the ABC test), it is the square root of the correlation of this independent OE test with the sum of the paired MC and OE tests.

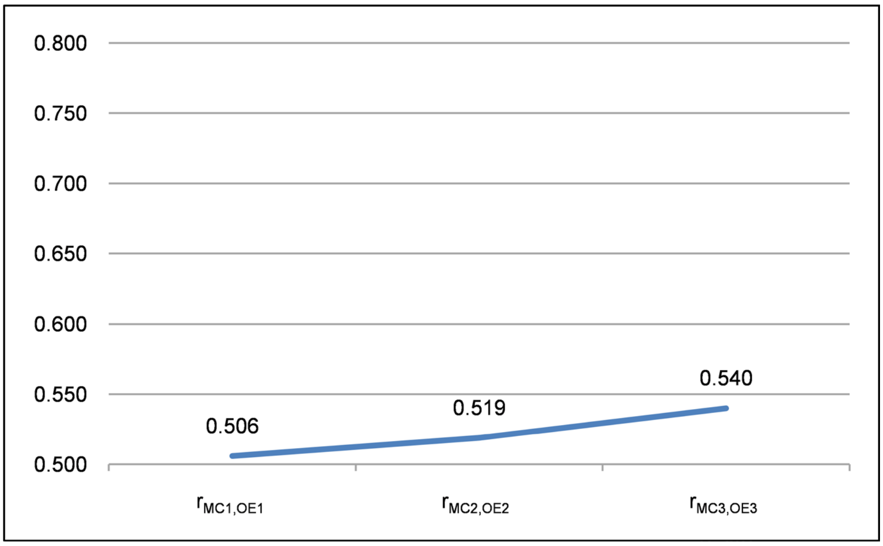

Figure 2. Increasing agreement indicated by small but significant increases (compare against Figure 3) in the correlations (plotted on the vertical axis) between increasingly complex and longer but very different test formats—MC (multiple-choice) and OE (open-ended) as plotted on the horizontal axis—assessing the same content as it accumulates over time in a course with 29 meetings over 16 weeks.

Table 2. Correlations and coefficients of determination (r^2) to test the experimental hypotheses pertaining to MC and OE tests.

Correlations of Single Concomitant Pairs of MC and OE Tests		Correlations of Pairs of MC and OE Tests 1 + 2, 1 + 3, and 2 + 3		Correlation of All MC Tests Combined with All OE Tests Combined	
$r_{MC1,OE1}$	0.506	$r_{1,2}$	0.544		
$r^2_{MC1,OE1}$	0.256	$r^2_{1,2}$	0.296	$r_{allM,allOE}$	0.792
$r_{MC2,OE2}$	0.519	$r_{1,3}$	0.641		
$r^2_{MC2,OE2}$	0.270	$r^2_{1,3}$	0.411		
$r_{MC3,OE3}$	0.540	$r_{2,3}$	0.769	$r^2_{allMC,OE}$	0.627
$r^2_{MC3,OE3}$	0.291	$r^2_{2,3}$	0.592		

Given that the lower bound of reliability on the 100 item MC tests averages above 0.90 and given that the lower bound on the reliability of any one of the OE tests cannot be less than the square root of the weakest of that time series of correlations—estimating that lower bound conservatively at 0.70 for each of the OE tests—to assess the significance of the growth in agreement from MC and OE pair 1, to pair 2, and then from 2 to 3, because the repeated-measures design orthogonalizes all variables except the time lapse and the growth in agreement from occasion 1 to 2 and 2 to 3, the numerator of the required F-ratio to assess the increase in r^2 from 1 to 2, and so forth, can be found by dividing the explained variance (the increase in r^2) by the geometric mean of the unreliabilities (conservatively estimated), the unexplained variance in the tests at issue, distributed over 77 minus 2 degrees of freedom:

$$F = \frac{r^2_{MC2,OE2} - r^2_{MC1,OE1}}{\sqrt{(1-0.90)*(1-0.70)}\big/(77-2)} \tag{1}$$

Hypothesis 1 must be accepted and the corresponding null must be rejected. Figure 2 summarizes the predicted and observed monotonic growth in agreement across the MC and OE pairs in the relevant time series. Compared against the increases seen in the middle column of Table 2 (displayed graphically in Figure 3), those in the first column of Table 2 (displayed graphically in Figure 2) are small (but nonetheless significant).

Hypothesis 2 can be tested against the progression of correlations in the middle of Table 2 (which contrasts are displayed graphically in Figure 3). As predicted, there are even larger gains in agreement across the time series when the respective pairs of MC and OE tests are cumulated so that each MC composite and each OE composite is doubled in length. Given the doubling of the length of the respective composites—according to the Spearman-Brown prophecy formula [50] which follows from the central limit theorem (generalized by Gnedenko and Kolmogorov, in 1968 [51]) and more particularly, in this instance, from the reduced error of the sampling distribution as the number of items is increased—reliability and validity of the composite tests should improve measurably as reflected in their variance overlap (agreement). The result shows that the increase in agreement variance from pair 1 to pair 2 is highly significant: F = 19.379, df 77, 2, $p < 0.0001$. Similarly, the increase in agreement from pair 2 to 3, calculated in the same manner, gives an F = 29.877, df 77, 2, $p < 0.0001$. Therefore, Hypothesis 1 must be accepted and the corresponding null must be rejected. Figure 2 summarizes the predicted and observed monotonic growth in agreement across the MC and OE pairs in the relevant time series. Compared against the increases seen in the middle column of Table 2 (discplayed graphically in Figure 3), those in the first column of Table 2 (displayed graphically in Figure 2) are small (but nonetheless significant).

Hypothesis 2 can be tested against the progression of correlations in the middle of Table 2 (which contrasts are displayed graphically in Figure 3). As predicted, there are even larger gains in agreement across the time series when the respective pairs of MC and OE tests are cumulated so that each MC composite and each OE composite is doubled in length. Given the doubling of the length of the respective composites—according to the Spearman-Brown prophecy formula [50] which follows from the central limit theorem (generalized by Gnedenko and Kolmogorov, in 1968 [27]) and more particularly, in this instance, from the reduced error of the sampling distribution as the number of items is increased—reliability and validity of the composite tests should improve measurably in variance overlap.

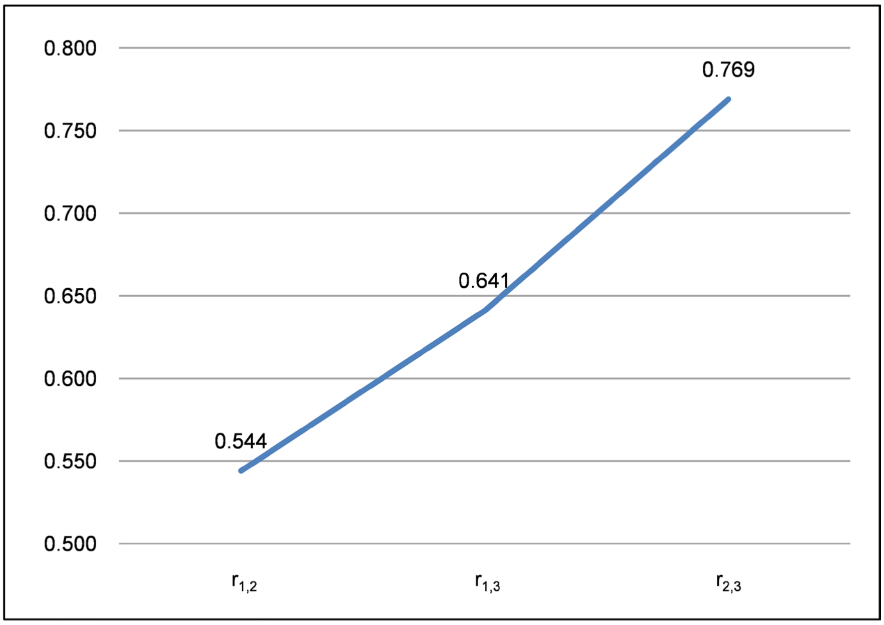

Figure 3. Increasing agreement indicated by substantial (compare with Figure 2) and highly significant increases in the correlations (plotted on the vertical axis) between MC (multiple-choice) and OE (open-ended) formats of increasing length and with increasing content accumulating over 29 meetings in 16 weeks.

The expected progression as summed up in Table 2 should therefore also be greater than the contrasts seen in Figure 2. By testing the least of the possible contrasts between the coefficients of determination (variance overlap) for the best performing single MC and OE pair 3 ($r^2_{MC3,OE3}$ = 0.291) against the worst performing composite where the MC tests in pairs 1 and 2 are correlated with the OE tests in those same pairs, $r^2_{1,2}$ = 0.296, and adjusting the estimates of reliability in the F-ratio denominator by the Spearman-Brown prophecy formula for doubling the length of the MC test from the 100 items in MC1 added to 100 items in MC2 gives a lower bound reliability estimate at 0.975 for the 200 item MC composite, and yields an estimated reliability of 0.737 for the least reliable of the 60 item OE composites, as seen in formula 2:

$$F = \frac{r^2_{MC3,OE3} - r^2_{1,2}}{\sqrt{(1-0.975)*(1-0.737)}\Big/(77-2)} \tag{2}$$

Working through the numbers yields an F-ratio of 30.377, df 77, 2, p < 0.0001 for the increase in agreement expressed in the numerator of formula 2.

Therefore, since all other possible contrasts between coefficients of determination for single concomitant pairs of MC and OE tests (the left-most column of Table 2), are greatly exceeded by all three of the coefficients of determination between the respective composite scores created by combining the MC parts of pairs 1 and 2 and the OE parts, as well as the composites for 2 with 3, and 1 with 3 (the middle column of Table 2), all of the contrasts are even more significant than the one examined. Therefore, every contrast in the increasing agreements/convergences predicted by Hypothesis 2 is confirmed and the corresponding null hypothesis, in all its parts, is rejected.

Finally, we come to Hypothesis 3. It can be tested straightforwardly from the right-most column of Table 2. Given the increase in length by summing 500 MC items, and summing the three OE tests based on study guide questions of the kind illustrated in MC and OE examples 1 - 3 above, plus the additional ABC OE test administered during class meeting 21, which adds 42 more OE items to the 90 from OE tests 1 - 3, again reliability of the composites can only be expected to be increased. However, without making any adjustments for increased reliability in the denominator of the appropriate F-ratio to test Hypothesis 3, the increase in agreement from the largest coefficient of determination for composite totals of pairs of MC and OE tests shown in the middle of Table 2 for $r^2_{1,3} = 0.592$ subtracted from the coefficient of determination for the correlation between all of the MC tests (500 items) and all of the OE tests (132 items) at $r^2_{\text{allMC, allOE}}$ = 0.627 produces an F-ratio for the increase in agreement at 206.843, df 77, 2, p < 0.000001.

Obviously, given the fact that all the other possible contrasts between coefficients of determination in the center column of Table 2 (also see Figure 3) are greater than the one just examined, all of them are significant as predicted by Hypothesis 3. Notably, based on the last composite correlation at 0.792, which must be read as a validity estimate for the agreement between MC and OE tests throughout the study period, the overall reliability of the OE composite score cannot be less than its lower bound at the square root of that correlation which is 0.890, a respectable value for any teacher-made test[2]. Using the correlation between pairs and composites of MC and OE test formats to measure agreement the degree of understanding of content measured by those disparate formats in a university course on human anatomy and physiology—the general thesis that better communication will produce better testing and teaching is sustained. The supposition that agreement should and does advance over the course of a semester of study as assessed by a powerful repeated-measures design is sustained. The increase in agreement across the diverse formats of test items is significant at each step along the way. Also, the reliability, validity, and interpretability of test scores is consequently enhanced by making them accessible to students throughout the course.

9. Value Added Measures of Success

Next we ascend to a higher level of abstraction to consider and test the theory behind "value-added" course/teacher evaluation. Edward Haertel [55] put the argument for it like this:

> It seems we hear daily about declining college and career readiness, 21st-century skills, and global competitiveness if public education does not improve… What could be more reasonable, then, than looking at students' test scores to determine whether or not their teachers are doing a good job? The teacher's job is to teach. Student test scores measure learning. If teachers are teaching, students should learn and scores should go up. If they are teaching well, scores should go up a lot. If test scores are not moving, then the teachers should be held accountable. (p. 14)

9.1. Aiming for Improvements

Seven successive iterations of the same course incorporated updated research findings, new tutorials, animations, demonstrations, PowerPoint slides, improved test questions, new theories and research findings. Although these changes should be expected to make the course more challenging the number of MC items from which tests were taken remained approximately the same across all iterations of the course. In the third edition a major innovation was added to the eBook. A way was discovered and implemented to create "loopy links" [56] connecting the MC and OE test items directly on a click with the relevant material in the eBook.

[2] What amounts to a respectable level of reliability is always a matter of subjective judgment, even among the professionals at or the Psychological Corporation where the largest standardized testing programs in the world are managed. At ETS Educational Testing Service (ETS), for example, [52] as well as [53], judged estimates for the Test of English for International Communication (TOEIC), a very widely used professionally prepared international test, at 0.79 to 0.86, to be "acceptably high" [54]. For teacher-made tests reliabilities are usually considered to be acceptable at much lower levels, but here, the overall reliability estimated for the MC and OE tests used in the courses at issue (0.89) exceeds reliabilities obtained for the TOEIC even in large scale testing at 0.79 to 0.80 [54].

A hyperlink going first to the item, and, then, looping back to the place where the student just left off reading was inserted for all of the 1087 test items in the third edition of the text. At the same time, making things more challenging for students, the correct answers to all of those items were no longer displayed in the item lists. They had to be worked out by the student[3]. It was estimated by the publisher that the amount of material was increased by approximately one-third with the publication of the third edition in the spring of 2020.

9.2. A Further Complication: COVID-19

Another complicating factor, was the forced move from face-to-face class meetings to one hundred percent remote teaching after March 16, 2020. That change was generally expected to reduce any gains underway because of any "value-added" from improved alignment of the subject-matter, methods of presentation, and testing. With that complication, additions to the subject-matter, and the additional burden put on students to construct answers for all 1087 test items, over the 7 iterations of the course, such increasing demands should bias things against the predicted Zolmogorov convergence and "added-value" from one semester to the next.

9.3. Results: "Added Value" for Seven Iterations

Figure 4 reports the results considering only MC test scores. The reason OE test items are not included in the comparisons of Figure 4 is because during semester 6, after the sudden shut-down of face-to-face class meetings during semester 5, instructors were urged by the university administration to reduce the load of required work. As a result, during semester 6, no OE tests were required and students were given the option of either working through the 19 plus hours of recorded lectures or merely reading the assigned material in the eBook, or both, as they might prefer, at their own pace, and on their own recognizance (no checking by the instructor other than scores on the MC tests). Also, students were permitted to repeat tests as many times as they might like although a different sampling from the respective pool of items up to the place of the test in the course sequence would be presented on each pass. The goal was to get students to address the conceptual content in all 1087 items in the pool. The only rigid constraint was that students had to finish all 9 MC tests (and whatever OE tests and other assignments were included) by the absolute deadline which was the end of the semester set by the university administration.

Having complied with the urging of the administration to make things easier in the spring and fall of 2020 (semester 5 and 6), in semester 7, I required a great deal more of the next group of students and expected to see a drop in scores. In the spring of 2021 (the 7th iteration of the course) I required completion of each

[3] However, with the reduction in travel to and from the campus, face-to-face conferences with students became common and immediate. Rarely would a question go unanswered for more than an hour. Direct contacts with students were faster, more effective, and far more frequent.

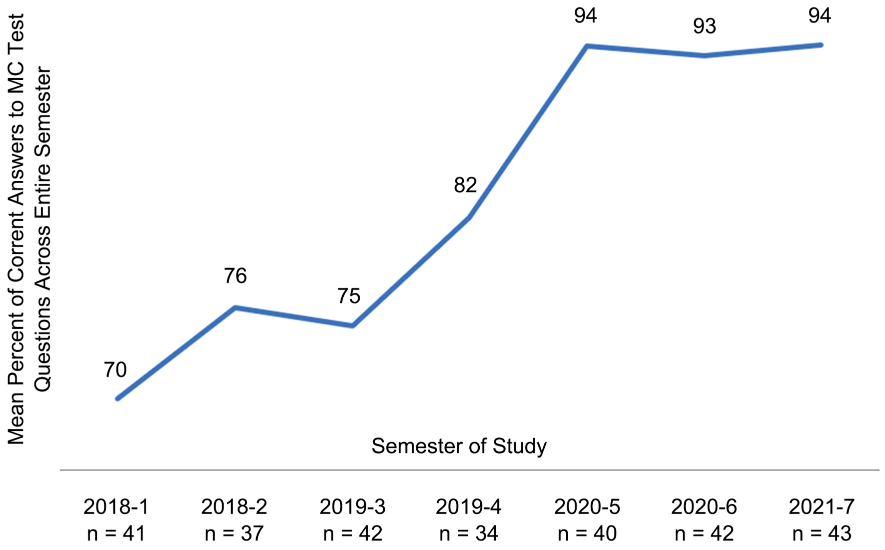

Figure 4. The mean score at the end of the semester on seven iterations of the same course on multiple-choice items cumulatively covering the whole subject-matter (increasing from 350 in semesters 1 to 3, to 620 in semester 7) in the same course, taught by the same teacher, with students in the same major field of study at the same university working through the same required curriculum ($N = 279$).

videographed lecture (approximately 19 and a half hours), the 9 OE tests, an essay for each of the 9 lectures, and completion of an essay-grading assignment to establish clearly what was wanted in each of the essays students wrote over each lecture. Instead of a drop from semester 6 to 7 I saw a slight improvement in spite of the fact that all the students were performing at near the ceiling on the tests. Given all the foregoing, the hoped for convergence toward unity seems to have occurred as the alignment between subject-matter, methods of presentation, and performance measures was adjusted across the various iterations of the course. Figure 4 shows value-added gains at about 24%.

10. Inevitable Conclusions

In all of the foregoing in PART ONE and in PART TWO, the not so simple contrast between *independent* and *dependent* sequences of events is central. The Kolmogorov-type of *independent* event sequences (e.g., the flipping of a coin—where knowing the outcome of any one event, or series of them, gives us exactly zero information about any of the rest) is profoundly different from deliberately constructed *dependent* sequences that intelligent instructors invent to share with cohorts of learners. Kolmogorov's two-page proof of "The Zero or One Law in the Theory of Probability", amplified especially here in PART TWO, shows that successful communications must outdistance failed efforts about as much as zero differs from unity. Failed experiments—supposing only that they are slightly intelligible and replicable—can never form an *independent* event sequence of the Kolmogorov type, but even if they could form such a sequence, no real string of them, no matter how long, would ever be sufficient to prove any null hypothesis.

There are multitudes of ways to fail to make a lightbulb that works, to end up short of climbing Mount Everest, and so on. Real successes are vanishingly few in comparison to possible ways to fail. For all those reasons, reliable and valid communications in instructional settings assessed with relevant and authentic real-life performances—intelligible, replicable, experimental outcomes—are perhaps not infinitely more useful than efforts that fail, but the difference favors successes by somewhere near the margin between 1 and 0.

Generalizable TNRs, and the theories derived from them, are the only valid basis for the kind of instruction that keeps scattering information to all takers and yet keeps on increasing. Shared information from valid teaching increases faster than we can give it away.

Acknowledgements

In addition to the people already acknowledged in the prior part, as the editor-in-chief of another open access journal, I salute those doing the editorial work at OALIB journal. I believe the future of scientific knowledge depends on open, preferably free, access to such knowledge. Any errors remaining in this paper are my own.

Conflicts of Interest

The author declares no conflicts of interest.

References

[1] Kolmogorov, A.N. (1933) Grundbegriffe der Wahrscheinlichkeitsrechnung [Foundations of the Theory of Probability]. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-49888-6

[2] Bayes, T. and Price, R. (1763) Essay towards Solving a Problem in the Doctrine of Chances. *Philosophical Transactions of the Royal Society*, **53**, 370-418. https://royalsocietypublishing.org/doi/pdf/10.1098/rstl.1763.0053?keytype2=tf_ipsecsha&ijkey=d86e9f6c361806fb58be6aad56cb2bcfade22c74

[3] Pólya, G. (1920) Über den zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung und das Momentenproblem [About the Central Limit Theorem of the Probability Calculation and the Moment Problem]. *Mathematische Zeitschrift*, **8**, 171-181. https://doi.org/10.1007/BF01206525

[4] Wikipedia (2021) Central Limit Theorem. https://en.wikipedia.org/wiki/Central_limit_theorem

[5] Kolmogorov, A.N. (1989) Publications of A. N. Kolmogorov. *The Annals of Probability*, **17**, 945-964. https://doi.org/10.1214/aop/1176991252

[6] Grunwald, P.D. and Vitanyi, P.M.B. (2003) Kolmogorov Complexity and Information Theory: With an Interpretation in Terms of Questions and Answers. *Journal of Logic, Language and Information*, **12**, 497-529. https://doi.org/10.1023/A:1025011119492

[7] Oller, J.W. (2010) The Antithesis of Entropy: Biosemiotic Communication from Genetics to Human Language with Special Emphasis on the Immune Systems. *Entropy*, **12**, 631-705. https://doi.org/10.3390/e12040631

[8] Oller, J.W. (2014) *Biosemiotic Entropy*: Concluding the Series. *Entropy*, **16**, 4060-

4087. https://doi.org/10.3390/e16074060

[9] Gryder, B., Nelson, C. and Shepard, S. (2013) Biosemiotic Entropy of the Genome: Mutations and Epigenetic Imbalances Resulting in Cancer. *Entropy*, **15**, 234-261. https://doi.org/10.3390/e15010234

[10] Oller, J.W., Shaw, C.A., Tomljenovic, L., Karanja, S.K., Ngare, W., Clement, F.M., *et al.* (2017) HCG Found in WHO Tetanus Vaccine in Kenya Raises Concern in the Developing World. *OALibJ*, **4**, Article No. e3937. https://doi.org/10.4236/oalib.1103937

[11] Oller, J.W., Shaw, C.A., Tomljenovic, L., Ngare, W., Karanja, S., Pillette, J., *et al.* (2020) Addendum to "hCG Found in Tetanus Vaccine": Examination of Alleged "Ethical Concerns" Based on False Claims by Certain of Our Critics. *International Jouurnal of Vaccine Theory*, *Practice*, *and Research*, **1**, 27-50.

[12] Reid, T.R. (1994) Airbus Crash Kills 261 in Japan. *Washington Post*. https://www.washingtonpost.com/archive/politics/1994/04/27/airbus-crash-kills-261-in-japan/ab814364-fd94-4ce4-a71f-5aaf223f7a4f/

[13] Day, B. (2004) Heightened Awareness of Communication Pitfalls Can Benefit Safety. *ICAO Journal*, **59**, No. 1, 20-22.

[14] Kolmogorov, A.N. (1965) Three Approaches to the Quantitative Definition of Information. *Problemy Peredachi Informatsii* [*Problems of Information Transmission*], **1**, 3-11.

[15] Kolmogorov, A.N. (1998) On Tables of Random Numbers. *Theoretical Computer Science*, **207**, 387-395. https://doi.org/10.1016/S0304-3975(98)00075-9

[16] Wikipedia (2021) Mutual Information.

[17] Oller, J.W. (1970) Transformational Theory and Pragmatics. *The Modern Language Journal*, **54**, 504-507. https://doi.org/10.1111/j.1540-4781.1970.tb03585.x

[18] Oller, J.W. (1971) Coding Information in Natural Languages. Vol. 123, De Gruyter Mouton, Berlin. https://doi.org/10.1515/9783111657219

[19] Oller, J.W. (1972) On the Relation between Syntax, Semantics and Pragmatics. *Linguistics*: *An International Review*, **10**, 43-54. https://doi.org/10.1515/ling.1972.10.83.43

[20] Chomsky, N.A. (1957) Syntactic Structures. De Gruyter Mouton, Berlin, Boston. https://doi.org/10.1515/9783112316009

[21] Berwick, R.C., Pietroski, P., Yankama, B. and Chomsky, N.A. (2011) Poverty of the Stimulus Revisited. *Cognitive Science*, **35**, 1207-1242. https://doi.org/10.1111/j.1551-6709.2011.01189.x

[22] Pinker, S. and Morey, A. (2014) The Language Instinct: How the Mind Creates Language. Unabridged Edition, Brilliance Audio, Grand Haven.

[23] Pinker, S. (2003) The Blank Slate: The Modern Denial of Human Nature. Reprint Edition, Penguin Books, London.

[24] Pinker, S. (2011) The Cognitive Revolution. *Harvard Gazette*.

[25] Sowa, J.F. (2018) Peirce, Polya, and Euclid: Integrating Logic, Heuristics, and Geometry. *Journal of Applied Logics*, **5**, 987-1059. http://www.collegepublications.co.uk/downloads/ifcolog00025.pdf

[26] Reichling, A. (1961) Principles and Methods of Syntax: Cryptanalytic Formalism. *Lingua*, **10**, 1-17. https://doi.org/10.1016/0024-3841(61)90108-5

[27] Uhlenbeck, E.M. (1967) Some Further Remarks on Transformational Grammar. *Lingua*, **17**, 263-316. https://doi.org/10.1016/0024-3841(67)90001-0

[28] Uhlenbeck, E.M. (1963) An Appraisal of Transformational Theory. *Lingua*, **12**, 1-18. https://doi.org/10.1016/0024-3841(63)90003-2

[29] Chomsky, N.A. (1965) Aspects of the Theory of Syntax. MIT Press, Cambridge, MA.

[30] Chomsky, N.A. (1980) On Cognitive Structures and Their Development: A Reply to Piaget. In: Piatelli-Palmarini, M., Ed., *Language and Learning*: *The Debate between Jean Piaget and Noam Chomsky*, Harvard University Press, Cambridge, 35-54.

[31] Berwick, R.C. and Chomsky, N.A. (2017) Why Only Us: Language and Evolution. Reprint Edition, The MIT Press, Cambridge, MA. https://doi.org/10.7551/mitpress/10684.001.0001

[32] Chomsky, N.A. (1995) The Minimalist Program. MIT Press, Cambridge, MA.

[33] Chomsky, N.A. (2011) The Machine, the Ghost, and the Limits of Understanding: Newton's Contributions to the Study of Mind. Center for the Study of Mind in Nature, University of Oslo, Oslo. https://www.reddit.com/r/philosophy/comments/3392qw/noam_chomsky_the_machine_the_ghost_and_the_limits/

[34] Oller, J.W. (1993) Reasons Why Some Methods Work. In: Oller, J.W., Ed., *Methods That Work*: *Ideas for Literacy and Language Teachers*, 2nd Edition, Heinle and Heinle Publishers, Boston, 374-385. https://www.researchgate.net/publication/235687222_Reasons_why_some_methods_work/stats#fullTextFileContent

[35] Peirce, C.S. (1897) The Logic of Relatives. *The Monist*, **7**, 161-217. https://doi.org/10.5840/monist18977231

[36] Peirce, C.S. (1909) Manuscript 514. In: Eisele, C., Ed., *New Elements of Mathematics*, Mouton, The Hague, 162-169.

[37] Peirce, C.S. (1902) Manuscript L75: Application to the Carnegie Institution (July 15, 1902) Analytical and Editorial Work by Joseph Ransdell. Ransdell J, Ed., Peirce Telecommunity Project: Electronic Peirce Consortium, Lubbock, Texas. https://link.springer.com/chapter/10.1007%2FBFb0027883?error=cookies_not_supported&code=f789ea72-3b27-46d1-9367-8147174af507

[38] Tarski, A. (1941) The Concept of Truth in Formalized Languages. *Journal of Symbolic Logic*, **6**, 73-89. https://doi.org/10.2307/2268577

[39] Tarski, A. (1949) The Semantic Conception of Truth. In: Feigl, H. and Sellars, W., Eds., *Readings in Philosophical Analysis*, Appleton, New York, 341-374.

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

[41] Peirce, C.S. (1865/1982) Harvard Lectures on the Logic of Science. In: Fisch, M., Kloesel, C.J.W., Moore, E.C., Roberts, D.D., Ziegler, L.A. and Atkinson, N.P., Eds, *Writings of Charles S Peirce*: *A Chronological Edition*, Vol. 1, 1857-1866, Peirce Edition Project, Indiana University Press, 1982, Bloomington, 256-257.

[42] Peirce, C.S. (1867/1982) The Logic Notebook. In: Fisch, M., Kloesel, C.J.W., Moore, E.C., Roberts, D.D., Ziegler, L.A. and Atkinson, N.P., Eds., *Writings of Charles S Peirce*: *A Chronological Edition*, Vol. 1, 1857-1866. Indiana University Press, Indianapolis, 344-357.

[43] Plothe, C. (2010) The Perinatal Application of Synthetic Oxytocin and Its Possible Influence on the Human Psyche and the Etiology of Autism. *Journal of Prenatal & Perinatal Psychology & Health*, **25**, 89-105. https://www.semanticscholar.org/paper/The-Perinatal-Application-of-Synthetic-Oxytocin-and-Plothe/800321a9be1575165ee90c4660f90773e3cc708c

[44] Sowa, J.F. (2017) Commentary on "Existential Graphs: MS 514 by Charles Sanders Peirce". MIT, Boston, MA. http://www.jfsowa.com/peirce/ms514.htm

[45] Sowa, J.F. (2017) Existential Graphs: The Simplest Notation for Logic Ever Invented. MIT, Boston, MA. http://www.jfsowa.com/talks/egintro.pdf

[46] Einstein, A. (1921) Geometry and Experience. Springer, Berlin, 1-10. https://www.relativitycalculator.com/pdfs/einstein_geometry_and_experience_1921.pdf

[47] Ryabko, B.Ya. (1986) Noiseless Coding of Combinatorial Sources, Hausdorff Dimension, and Kolmogorov Complexity. *Problemy Peredachi Informatsii [Problems of Information Transmission*, **22**, 170-179. https://link.springer.com/chapter/10.1007%2F3-540-51498-8_42

[48] Kieffer, J.C. and Yang, E. (1996) Sequential Codes, Lossless Compression of Individual Sequences, and Kolmogorov Complexity. *IEEE Transactions on Information Theory*, **42**, 29-39. https://doi.org/10.1109/18.481775

[49] Maxwell, S.E. and Delaney, H.D. (1990) Designing Experiments and Analyzing Data: A Model Comparison Perspective. Wadsworth, Belmont.

[50] Wikipedia (2021) Spearman-Brown Prediction Formula. https://en.wikipedia.org/wiki/Spearman%E2%80%93Brown_prediction_formula

[51] Gnedenko, B.V. and Kolmogorov, A.N. (1968) Limit Distributions for Sums of Independent Random Variables. Addison-Wesley, Boston. https://www.amazon.com/Limit-Distributions-Independent-Random-Variables/dp/0201024209

[52] Liao, C.-W. and Qu, Y. (2010) Alternate Test Forms Test-Retest Reliability for the TOEIC Speaking and Writing Tests. In: Powers, D., Ed., *Research Foundation for TOEIC*: *A Compendium of Studies*, Educational Testing Service, Princeton, 11.1-11.40. https://www.ets.org/research/policy_research_reports/publications/report/2010/itka

[53] Liao, C.-W. and Wei, Y. (2010) Statistical Analyses for the TOEIC Speaking and Writing Pilot Study. In: Powers, D., Ed., *Research Foundation for TOEIC*: *A Compendium of Studies*, Educational Testing Service, Princeton, 9.1-9.25. https://www.ets.org/research/policy_research_reports/publications/report/2010/itjz

[54] Qu, Y., Schmidgall, J., Cid, J. and Chan, E. (2019) Linking OPIc Levels to TOEIC®Speaking Scores. Educational Testing Service, Princeton. https://www.ets.org/Media/Research/pdf/RM-19-02.pdf

[55] 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.

[56] Oller, J.W. (2020) How to Write Valid Test Items for Online Teaching and Testing: Games Students Can Play and Win. https://www.youtube.com/watch?v=kYdAQLgPZrk