RANDOMLY PARALLEL TESTS

AN ANNOTATED BIBLIOGRAPHY

The concept of randomly parallel tests (RPTs) was formally introduced by Frederic Lord in 1955. He defined RPTs simply as tests “consisting of a random sample of items drawn from a common population of items…” (Lord, 1955, p. 1). He continued supporting RPTs throughout his career at the Educational Testing Service, including providing theoretical and statistical justification, along with analysis tools to help the practitioner. Here are the papers I have found by Lord and others, including one of mine, that speak to this presently overlooked test design, a design that can solve many of the difficult problems in testing that we experience today, including the damage caused by cheating and test theft.

The citations are given in chronological order as accurately as possible based on the date of publication.

The annotations are my opinion about the content I read in the papers. Other topics besides RPTs and associated theory and procedures are discussed. I expect readers to discover more interesting information, hopefully supporting the use of RPTs today.

DAVID FOSTER | CAVEON

RPT REFERENCES

Lord, F. M. (1955). Sampling fluctuations resulting from the sampling of test items. Psychometrika, 20(1), 1-22.

In this paper Lord first introduces the concept of RPTs (p. 1) and also the terms, "randomly parallel forms or randomly parallel tests." He compares the process with the sampling that occurs when a random sample of examinees is selected for a research study. He proposes a method for calculating the standard errors for individual test scores, and presents a formula for test reliability, the KR21. He also describes the shape of the distribution of standard errors as binomial.

DFF Note: This paper is actually a revised version of an earlier, December 1953, paper by Lord, provided as a technical report to the Office of Naval Research based on a contract the ONR had with Educational Testing Service. The paper was titled, The Standard Errors of Various Test statistics When the Test Items are Sampled. Therefore, December 1953, is the actual date for the introduction of Randomly Parallel Tests.

DFF Note: In a few papers, Lord used different terms for tests that were "constructed" to be "parallel." He called these rationally parallel, statistically parallel, nominally parallel, and strictly parallel. These terms seem to be interchangeable.

Lord, F. M. (1955). Estimating test reliability. Educational and psychological measurement, 15(4), 325-336.

Lord discusses different issues for persons attempting to estimate test reliability, including the assumptions underlying different statistics, and the importance of the "exact definition of 'parallel'" (p. 325). To elaborate, he proposes two new definitions of parallel test forms, RPTs, and Stratified RPTs (Lord referred to the latter as "matched-forms" tests in this paper), defined as stratifying the item pool on one or more characteristics of the items in advance of random sampling. He suggests that these two new definitions of parallelism seem "to have at least as good justification as those usually used, and perhaps better." (p. 325).

DFF Note: We need to remember that the overall testing context for Lord's writing was, for the most part, traditional paper-and-pencil testing, and it had been for 40 years. Computers were talked about in the 1950s, but as a future technology, mostly. It would have been impractical to build RPTs at that time given the available technology and dominant paper-and-pencil testing approach. Without computers, randomization and having "large pools" of items were simply not feasible for most testing settings or for testing on a large scale.

Lord, F. M. (1959). Randomly parallel tests and Lyerly's basic assumption for the Kuder-Richardson formula (21). Psychometrika, 24(2), 175-177.

This is a short note clarifying the assumptions behind the use of RPTs and the use of the KR-21 reliability statistic.

DFF Note: It is worth careful reading as it deals with "item equivalence" and provides insight into how the number of items for a RPT affects the comparability of the scores from them: the more items in the RPT, the greater the ability to compare the scores.

Lord, F. M. (1959). An approach to mental test theory. Psychometrika, 24(4), 283-302.

Here, Lord describes several models, two of which describe the Stratified-RPT and the RPT as he has talked about in earlier articles. Each model is evaluated in terms of its assumptions, its definitions regarding true scores, the distributions of observed scores, and the distributions of measurement errors.

Lord, F. M. (1959). Tests of the same length do have the same standard error of measurement. Educational and Psychological Measurement, 19(2), 233-239.

This is an impressive discussion about standard errors of measurement. Lord provides empirical evidence from data from a wide variety of exams analyzed at some point by ETS scientists. The result was the conclusion stated in the title, but which the paper made clear applied to single individuals who had been or would be administered RPTs, not "rationally equivalent" tests.

DFF Notes: In several papers, Lord takes care to point out the mistaken conclusion that a test has a single standard error of measurement. He often points out that standard error of measurement is different for each test score.

Lord, F. M. (1959). Statistical inferences about true scores. Psychometrika, 24(1), 1-17.

Lord describes the benefits of applying the principles of random sampling of items from a large pool (creating RPTs by Type-2 sampling), to the example of matrix sampling, adding that examinees can also be considered as being randomly sampled from a population of examinees (Type-1 sampling).

Webster, H. (1960). A generalization of Kuder-Richardson reliability formula 21. Educational and Psychological Measurement, 20(1), 131-138.

Webster supports Lord's notion of randomly parallel tests and provides some advice on using the KR-21 as a measure of reliability for RPTs.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. British Journal of Statistical Psychology, 16(2), 137-163.

In comparing approaches to calculating reliability, Cronbach and his co-authors describe the advantages of assuming that items are randomly sampled. They state, "Randomness of sampling guarantees that in the population the means, variances, and intercorrelations of scores will be equal. One set of such randomly generated data is equivalent to another even though the tests individually are not equivalent." (p. 143)

DFF Note: The authors point out only two objections to the random-sampling model. The first is that universes, like domains, are usually vaguely defined. Second, that strict random sampling never occurs in practice.

Lord, F. M. (1964). Nominally and rigorously parallel test forms. Psychometrika, 29(4), 335-345.

This paper does not reference RPTs, but goes into detail on two alternatives, Nominally Parallel Tests and Rigorously Parallel Tests.

DFF Note: This paper is included in this bibliography to help place Lord's recommendations of RPTs in a broader context. Considering RPTs as a model for test construction and use is a choice that psychometricians have. That choice relies on comparing practical, statistical, theoretical advantages, and disadvantages.

Rajaratnam, N., Cronbach, L. J., & Gleser, G. C. (1965). Generalizability of stratified-parallel tests. Psychometrika, Vol. 30(1), March 1965.

In this paper the authors contrast the equivalence of test forms supported by classical theory with the item sampling models (with and without stratified random sampling). A part of the introductory paragraph provides the rationale for their work: "An alternative model (DFF Note: the term "alternative," referring to the rationally equivalent forms or equivalent-composites model from classical theory) which has received increasing attention in recent years regards a given measure as a random sample from a universe of measures whose homogeneity or equivalence is not specified a priori, and a composite test as a random sample of items from a universe of not-necessarily-equivalent items." (p.39) The authors add a third model, describing the stratified random sampling model. Given the interest of Cronbach and his colleagues in Generalizability Theory, an important outcome of using RPTs or Stratified-RPTs is the enhanced ability to generalize from a test score to the content universe or content domain of interest.

DFF Note: These authors, as well as others in the middle of last century, dealing with the topic of obtaining actual random samples, generally acknowledge the difficulty, and even the impossibility at that time, of creating a practically useful test by randomly sampling items. While randomly sampling items is a better model is a better model for testing in almost every sense, the practical use in operational testing is a significant barrier. Often proposed is an interim assumption and rationale that the items could be considered to have been randomly sampled from a universe or population of items (see p. 43). This paper and others use logic of this sort to circumvent the barrier. Of course, with the technology of the 21st century, pure RPTs are not difficult at all to create.

Lord, F. M. (1965). Item sampling in test theory and in research design. ETS Research Bulletin Series, 1965(2), i-39.

In this paper Lord uses the model term for RPTs, "the item sampling model." The description of the item sampling model is where the number of items on a test form "are considered as a random sample from a population of items." (p.1) He writes about the simplicity of the model, its minimal assumptions, and that it "yields many important results." (p.1) How to derive those important results is the purpose of the paper. He acknowledges a common objection to the item sampling model as simply that sampling items from a population of items is not ordinarily done. Quoting from p.4, Lord states, "In line with this reasoning (DFF Note: that a random sample is the best kind of sample to have), in testing work it will sometimes be essential to actually select items at random. In certain situations, this will be the only way to secure a firm basis for the necessary statistical significance tests and statistical inferences." The paper contains much more statistical reasoning, and many more practical insights into errors of measurement and true scores.

Lord, F. M., Novick, M. R., & Birnbaum, A. (1968). Statistical theories of mental test scores.
Lord, F. M., & Novick, M. R. (2008). Statistical theories of mental test scores. IAP.

Specifically Chapter 11, titled, Item Sampling in Test Theory and in Research Design.

Touted as one of the most important books in the history of psychometrics, this book was originally published in 1968, and republished in 2008. The chapter of interest that covers item sampling, or random sampling of items for tests, is Chapter 11, titled Item Sampling in Test Theory and in Research Design. It is devoted entirely to the case of a test where "n items are considered as a random sample from a population of items." (p. 234) The chapter makes a strong case for the use of RPTs in a wide range of testing circumstances.

DFF Notes: Lord and the other authors provide an interesting footnote on the first page of this chapter. It reads, "Reading of this chapter can be omitted without loss of continuity." To me, this suggestion highlights the unique nature of RPTs, which in 1968 didn't mesh well with classical theories, and which were still not feasible as part of an operational testing program, whether for large-scale or small-scale testing. However, it is clear that the authors believe it deserved an important place in discussions of theory and practice. They may have expected that with the coming of computerization of testing, it would be useful to researchers and practitioners sooner rather than later.

Osburn, H. G. (1968). Item sampling for achievement testing. Educational and Psychological Measurement, 28(1), 95-104.

Osburn proposes a way to build RPTs or Stratified RPTs using item forms (DFF: similar to AIG item models or Caveon's SmartItems™) with the goal of obtaining test scores that generalize to a universe or domain. He points out the importance of describing well the universe or domain.

Prosser, F., & Jensen, D. D. (1971, May). Computer generated repeatable tests. In Proceedings of the May 18-20, 1971, spring joint computer conference (pp. 295-301).

These authors describe several problems of traditional testing in higher education. In that context they then recommend using computers and Stratified-RPTs to create paper tests that are unique, that can be administered more frequently, provide immediate feedback,

and are repeatable. Repeatable, in this paper, refers to a process whereby students can take unique just-printed test forms on a course topic as often as desired to meet instructional goals. The items are created in advance, about 6x to 10x the number of items needed for any particular student's test form. The paper did not provide any research data for the recommended process actually used for a university course.

DFF Note: Creating unlimited and unique repeatable tests is viewed by these authors as well as Lord (1977) himself as a significant benefit of using RPTs or Stratified RPTs. The concept of test forms being unique and equivalent (parallel), and which can be created easily by a computer, should be highly attractive in any instructional setting as it removes the typical restriction of synchronous use for traditional tests. Unlimited, repeatable tests would be valuable in all areas of high-stakes or low-stakes testing.

Cronbach, L. J., Gleser, G. C., Nanda, H., and Rajaratnam, N. (1972). The dependability of behavioral measurements. Theory of generalizability for scores and profiles, 1-33.

Cronbach is creating and justifying Generalizability Theory with this book. In Chapter 11, titled Contributions and Controversy – A Summing Up, RPTs are mentioned as coming from Lord's work, and that they share similar assumptions of random sampling to his (Cronbach's) proposals (see p. 357). This reference is included also because it details some criticisms of the process of randomly selecting items for a test from a universe or population of items, including some concerns from R. L. Thorndike. Those arguments are presented and commented on by Cronbach, et al. These criticisms and comments are found mainly from pages 376 to 383.

DFF Notes: I often wonder why RPTs were never viable, at least until now, as a part of operational testing programs. No doubt the unavailability of computer technology played a role, but it may also may be a factor that such criticisms from proponents of traditional testing approaches inhibited testing professionals and other practitioners from changing their test designs away from something that had been "traditional" for decades.

Millman, J. (1973). Passing scores and test lengths for domain-referenced measures. Review of Educational Research, 43(2), 205-216.

Millman's paper provides practical advice for anyone desiring to create tests where resulting scores indicate the proportion mastered of a domain of content (represented by a large number of items). Random sampling of items from the pool of items can be

considered RPTs as Lord has described them. Millman discusses how mastery scores can be derived and how test lengths can be determined using the RPT model. In this paper Millman also describes "sequential testing" as computerized testing where the number of items in the test is continuously monitored and the scores compared with a passing standard.

DFF Note: Millman is creating a bridge between RPTs, bringing along their statistical and theoretical advantages, to the growing (at that time) field of domain-referenced testing. His paper also presents an early benefit of using computers for test administration.

Emerson, P. L. (1974). Experience with computer generation and scoring of tests for a large class. Educational and Psychological Measurement, 34(3), 703-709.

This paper describes a RPT where the item pool consisted of about 500 items (about 50 per chapter of a textbook) supplied by a textbook publisher. A computer system generated chapter tests by randomly selecting 20 items from the chapter-based strata of the pool. In some circumstances, students were given the opportunity to re-test. Unique final exam forms of 50 items were created using a Stratified RPT procedure (5 items randomly selected from each chapter). Even in 1974, the author concluded that the costs associated with this process was no more than they would have been for a conventional testing of course learning, and he concluded with recommending the process to other instructors.

Lord, F. M. (1977). Some item analysis and test theory for a system of computer-assisted test construction for individualized instruction. Applied Psychological Measurement, 1(3), 447-455.

Following up on the advantages of unlimited repeatable tests from RPTs, and with computers becoming more viable in the test development and test administration efforts, Lord provides direction as to the analysis of item statistics, test reliability, test scores and standard errors. In one data set he presented he showed that standard errors were lower for RPTs, explaining that "the pool of items is much better represented when each examinee takes a different set of 41 items than when one set of 41 items is used for everyone." (p. 454) Lord also estimates that the number of items in the pool should be 10x to 40x the number of items chosen for any test form.

DFF: As computers become more common and useful, Lord as well as others are describing more situations where using RPTs will be practical and beneficial. The reader should keep in mind that the perspective of these theorists and researchers is firmly

based in test administration that is paper-based and where item pools are formed in advance. It isn't until a few years later that computers are viewed as a way to actually administer exams, enabling concepts such as computerized adaptive testing, LOFTs, items that can be created when needed, and even SmartItems that can be used to render "items" on the fly during an exam.

Millman, J. (1977). Creating Domain-Referenced Tests by Computer. Paper presented at the Annual Meeting of the American Educational Research Association, New York, April 4-8, 1977.

This paper seems to be an earlier version of Millman & Outlaw (1978). However, it contains a description of how Millman created 132 "item programs" that serviced 7 RPT mastery tests for a course on statistics. The paper also reveals tips on how to evaluate the quality of the items produced on-demand by the item program.

Millman, J., & Outlaw, W. S. (1978). Testing by computer. AEDS Journal, 11(3), 57-72.

The authors describe using computers to create RPTs. As a unique feature, the study stored "item programs" rather than actual items in the system. The items themselves were generated for the students' unique tests just prior to the tests being printed and administered in paper-and-pencil form. This is a different approach to RPTs, using the random elements of item programs to create tests, rather than creating a large pool of items from which the items were randomly selected. Advantages of RPTs listed are familiar by now, including repeatable tests, providing low-cost and faithful practice assessments, control over cheating, providing make-up tests, and others.

DFF Note: The concept of "item programs" from this paper is similar to AIG item models and SmartItems™. AIG item models are built to create items for security reasons, to use in traditional tests. This means that the items automatically generated are stored in item banks and follow the same procedures as SME-created items to qualify them for use on operational tests. SmartItems are used directly on tests and render versions of themselves (called renderings, and sometimes, items) in real time. These renderings never see the inside of an item bank and are not "stored" except for research and legal purposes. Quality assurance steps are taken to make sure that SmartItems create item renderings on-the-fly that are of consistent and sufficient quality.

Cronbach, L. J. (1990). Essentials of Psychological Testing. New York, NY: Harper Collins Publishers.

This is the 5th and final edition of a popular textbook on principles of psychological measurement. Cronbach describes RPTs in Chapter 2, in the section titled, Testing in the Computer Age, mainly on pages 46 and 47. He begins the section with a statement about standardization: “Because of its consistency, the computer carries standardization to an extreme, yet it can achieve standardized measurement while presenting different questions (and personalized feedback) to every test taker.” (p. 46). Then he goes on to describe a context of testing for civil service jobs, and that unique tests can be built for each job candidate. He goes on to suggest that a computerized item model or item program could “arbitrarily” (e.g., on the fly during an exam) alter rates, make-up of shipments, and rules, adding a comment that the test taker gains no advantage from knowing the content of questions presented to a friend who had tested a few days earlier.

DFF Notes: Cronbach clearly provides, in 1990, a useful and refreshing view of standardization in the coming computer age, describing how the process might work in a manner similar to Lord’s RPTs or Caveon’s SmartItems™, along with how it would prevent cheating by pre-knowledge.

SOME DFF NOTES ON CURRENT AND POSSIBLE TECHNOLOGY-BASED RPT APPROACHES

Foster, D. F., The SmartItem (2020). https://info.caveon.com/the-smartitem-ebook-promo

A couple of years ago I wrote a booklet about SmartItems™. At the time, and still today, it was about providing unique exams to each test taker with the goal of making theft of exams, along with other forms of cheating, impossible. SmartItems are programmed to cover a content or skill domain in breadth and depth, or to cover a stratum of a domain. Caveon’s testing system, Scorpion, facilitates the development of SmartItems™ in several different ways, and also provides the means to administer tests comprised of SmartItems™. It wasn’t until after this book was released that I became aware of Lord’s writing and that SmartItems™ fit easily into his description of RPTs. I’m writing the 2nd edition, which will switch the focus from SmartItems™ to the broader concept of RPTs. SmartItems™ can be considered to be a reasonable way to use technology to implement RPTs.

Marder, A. A SmartItem Simulation (2019). https://amarder.shinyapps.io/smartitems_simulation/

This is a set of simulations of SmartItems, and, therefore, RPTs. All of the simulations compare SmartItems™ (or RPTs) with a test form with fixed items. The basic simulation allows you to vary the number of items on a test, the number of test takers, and the range of difficulty of the renderings from SmartItems™. (Some SmartItems™ can cover a narrow domain where there would be less variability for renderings; others would cover broader domains with a greater range of difficulty in the renderings.) Some test statistics, charts comparing estimated and true ability, and test information curves are produced. The second simulation, labeled Basic: Length, explores the effects of test length on reliability and error (median absolute deviation). The third and fourth simulations are similar to the first two, but allow for variation of item exposure and test taker pre-knowledge, and measure the effectiveness of pre-knowledge on the output variables. In general, the simulations indicate good comparability of test statistics between tests using SmartItems™ and fixed-item tests. It also shows the poorer performance of fixed-item tests, compared to SmartItems™, with different pre-knowledge conditions. A simulation similar to this, but based on the logic, theory, and analyses for Lord's RPTs, is currently in production.

DFF Note on LOFTs and RPTs

A LOFT or Linear-On-the-Fly Test is a computerized exam that is built during an exam sitting by drawing items randomly (by strata or not) from a pool, usually a pool that is stratified by content, item statistics or parameters, exposure rates, etc. In that sense, LOFTs are exactly what Lord described as RPTs in his initial papers written in 1955. It's the earliest description of LOFT that I have read. I have not seen people referring to LOFT who also credit Lord for its genesis, so that Lord invented LOFT might not be well known.

The only difference that I can see between Lord's description of LOFT and how LOFT is described and used today is that Lord qualified that the pool had to be "large," providing the range of 10x to 40x the number of items in the test. He made this recommendation for at least two reasons. First, test security. Providing unique tests to individuals would make cheating more difficult. His second reason is repeatability. Having more items would allow the test to be repeatable, even for an individual. This would be helpful in educational settings where the test could be used to measure a student's learning before, during and after instruction. Today's LOFTs would not qualify as RPTs because the pool is too small. I would recommend increasing the size of the pool to meet Lord's guidelines.

DFF Note on AIG and RPTs

AIG, or Automated Item Generation, is becoming more popular as a way to increase the

AIG, or Automated Item Generation, is becoming more popular as a way to increase the number of items for a testing program, mainly to support security activities, such as the creation of more equivalent forms, increasing the size of a CAT pool, or replacing compromised items in existing operational exams. AIG can be generally characterized as the development of item models or templates designed to generate items automatically that are appropriate for the tests an organization wishes to build. An item model is like a manufacturing assembly line, combining components of items with data sources and according to rules, with the result being new, useful items, perhaps thousands or tens of thousands of them. Items produced using AIG are stored in item banks, and may need to undergo additional quality steps, such as field testing or expert reviews.

It's not much of a leap to see that an AIG process could create every item for the large pool of items Lord envisioned for RPTs.

There is perhaps another use of item models than the one described above. Once vetted for the quality of item production, perhaps each item model could serve on an operational exam to produce items in real time as needed by individual test takers. This is similar to how SmartItems are used.

DFF Note on CATs, RPTs and SmartItems

CATs already produce relatively unique tests for examinees. The larger the pool of items supporting the CAT, the greater likelihood that individual exams will be unique. A smaller size of pool results in more overlap of items across test takers. A lot of overlap encourages harvesting and sharing of items, and contributes to the effectiveness of cheating on CATs. An early scandal in one of the first uses CATs exposed this problem of small pools.

CATs can be considered a variation of RPTs where the items are randomly drawn from strata organized at least by difficulty, but also by exposure rate and content. Another difference is that test taker ability is a major determiner of which items are selected from the pool. Random selection is used as part of many CAT selection algorithms to prevent the higher quality (most informative) items from being used too often.

Lord's rules for the size of LOFT pools of 10x to 40x would apply equally to CATs. Adjusting those multipliers for today's security threats and technology would suggest a pool size of 100x or even more.

SmartItems are defined as actual items. Within a content domain, some SmartItems would be more difficult than others. Statistical calibrations would provide the IRT parameters needed for the CAT. As an example, if the four primary mathematical operations were the domain, then a SmartItem producing addition renderings would likely be easier overall than a SmartItem producing multiplication renderings. In that case, those two SmartItems and others could be combined in a CAT item pool to support the measurement of a young student's abilities in mathematics operations.