2014

Beneficial Assessment Outcomes from Frequent Testing

Abdulrazaq Imam
John Carroll University, aimam@jcu.edu

VOLUME 20 ISSUE 2

The International Journal of

Assessment and Evaluation

Beneficial Assessment Outcomes from Frequent Testing

ABDULRAZAQ A. IMAM

COMMON GROUND

THELEARNER.COM

THE INTERNATIONAL JOURNAL OF ASSESSMENT AND EVALUATION
www.thelearner.com

First published in 2014 in Champaign, Illinois, USA
by Common Ground Publishing LLC
www.commongroundpublishing.com

ISSN: 2327-7920

The International Journal of Learning in Assessment and Evaluation is peer-reviewed, supported by rigorous processes of criterion-referenced article ranking and qualitative commentary, ensuring that only intellectual work of the greatest substance and highest significance is published.

Beneficial Assessment Outcomes from Frequent Testing

Abdulrazaq A. Imam, John Carroll University, U.S.A.

Abstract: When faced with deadlines, people tend to procrastinate. Students do this by delaying study time until examinations are so close the only option left is cramming. This procrastination scallop is a well-established behavioral phenomenon in both human and infrahuman species. Distributed practice also has been demonstrated to be superior to massed practice in the cognitive literature. Frequent testing provides opportunities for distributed practice and rehearsals that fill the gap between acquisition and the big test, creating its own mini-scallops. In sections of Introductory Psychology, Research Design, and Learning and Behavior courses, standard pre-post testing was conducted at the start and end of the semester over many years. No weekly quizzes were required in one course for a few semesters, in contrast to the remaining courses. Mean assessment gains were substantially bigger with than without weekly quizzes and the difference was statistically significant. The results indicate beneficial assessment gains in learning from frequent quizzes and suggest potential alternative strategies for faculty to implement low-cost effective instructional practices that students may benefit from

Keywords: Pedagogy, Quantitative Assessment Outcomes, Methodology

To test or not to test? Not testing usually is not an option in today's education environment. We have to do it, but how often? At an absolute minimum, one must test at least once per term. In common practice, however, there is often at least a midterm and a final examination. How much testing is pedagogically sound for effective learning? Some give tests every month of the semester. Is there justification for limiting or expanding how much testing occurs in a typical college course? On the one hand, students may not be fond of more testing because it means studying more over the term. On the other hand, the instructor may see offering more tests than the typical midterm and final as more work, preparing the additional tests and grading them. There are, however, behavioral and cognitive rationales for recommending more frequent testing than is usual in college courses. Indeed, in a recent issue of Psychological Science in the Public Interest, Dunlosky, Rawson, Marsh, Nathan, and Willingham (2013) highlighted a number of learning techniques that students may avail themselves of for better learning outcomes. In what follows, I present some behavioral and cognitive principles that provide justification for more frequent testing than is typically used in college courses today.

Behavioral Rationales

When faced with fixed, predictable deadlines, we tend to procrastinate. In the laboratory, animals show this in the form of fixed-interval (FI) pause and run performance shown in cumulative records of their responses (see Figure 1). In this case, when food is delivered every 5 min (FI 5-min schedule), the rat pauses immediately following the last food delivery and only begins to respond when the next delivery is eminent. This pattern of pause and run is what is described as scalloping (Ferster and Skinner 1957). Although human FI performance may manifest scallops under special conditions in the laboratory (e.g., Weiner 1962; see Wanchisen, Tatham, and Mooney 1989, for a nonhuman example), procrastination is ubiquitous in the real world. Even the United States Congress has been shown to exhibit this procrastination pattern (Critchfield, Haley, Sabo, Colbert, and Macropoulis 2003). Students' study behaviors are not exempt from the procrastination scallop (Michael 1993). Michael provides a schematic of various scallops for study behavior reflecting different points from inception of tasks to their completion, showing "safe" procrastination periods during which students tend not to study and depicting their implications for exam and course outcomes. The scallops show that students tend to wait early during the period and resume studies only as the next exam approaches. If they wait too long, as many do, only very little time is left and they encounter the skull, cramming as much information

in as possible, but then end up doing poorly on the exam. If they start early during the "safe" period, they tend to do better on the exam; those who start earliest do the best. The problem is how do you get students to start early?

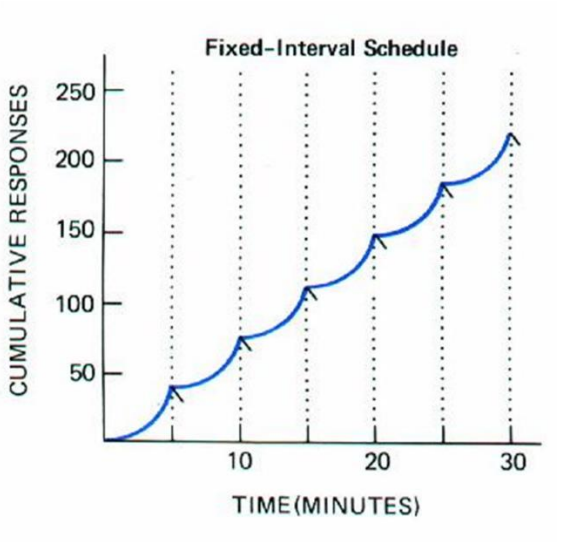

Figure 1: Cumulative record of responding on a FI 5-min schedule of reinforcement.

There is some evidence in the literature that more frequent testing prevents procrastination (Mawhinney, Bostow, Laws, Blumenfeld, and Hopkins 1971). Mawhinney et al. conducted daily versus three-week testing in their second experiment and recorded the minutes spent studying during study sessions. When subjects were tested daily, they studied more consistently than when tested every three weeks; under the three-week testing condition, they tended to study less early compared to close to the test, thereby producing scallops. Notably, subjects studied longer than their daily average level immediately just before the test in the three-week condition having studied less early during the period, suggestive of cramming behavior. Cramming, of course, represents massed practice, which usually results from lack of preparation due to procrastination.

Cognitive Rationales

Massed practice and distributed practice have received extensive attention in the cognitive literature (see Cepeda, Pashler, Vul, Wixted, and Rohrer 2006, for a recent review). The general finding has been that distributed practice is superior to massed practice (Dunlosky et al. 2013; Roediger 2013). Bahrick (1979) is illustrative: For long-term retention, the more effective approach for students is distributed practice. Bahrick compared recall of Spanish translations following six practice sessions with 0-, 1-, or 30-day delays between sessions. With 0-day between practices, subjects scored higher on Spanish translations compared to those using 30-day between practices on tests just before the practice session. What is interesting, however, is how well the latter subjects did on a test 30 days after the practice sessions when they outperformed those with 0- or 1-day between sessions (see Dunlosky et al. 2013); the difference is in distributed versus massed practice, respectively. As Roediger noted, "[l]earning can occur quickly under massed-practice conditions, so it seems like an efficient way to teach, but hundreds of studies have shown that distributed practice leads to more durable learning" (2013, 3).

What we know about memory processes in terms of levels of processing (Craik 2002; Craik and Lockhart 1972; Moscovitch and Craik 1976) suggest that students should engage in elaborative rather than maintenance rehearsals in order to properly encode information into long-term memory. Better, longer retention tends to follow meaningful encoding of information, perhaps requiring some "consolidation" (Craik 2002, 310), which Cepeda et al.'s (2006) review suggests may account for distributed practice effects. Due to the nature of massed practice in cramming for examinations, meaningful encoding or elaborative rehearsal is precluded by the need to get in as much information as is possible in a very limited time. In contrast, with distributed practice, time availability between practices allows for the type of integration that might be needed to achieve longer-term retention of material. In this way, distributed practice is especially conducive to elaborative rehearsals and thus long-term retention. What weekly quizzes do is provide for distributed practice as well as opportunities for rehearsal of covered material before the big test, thereby preventing cramming. Figure 2 shows how such testing opportunities create interim mini scallops that culminate in the big exam, thereby thwarting procrastination. The question that remains is how effective is the use of weekly quizzes and how to demonstrate that effectiveness.

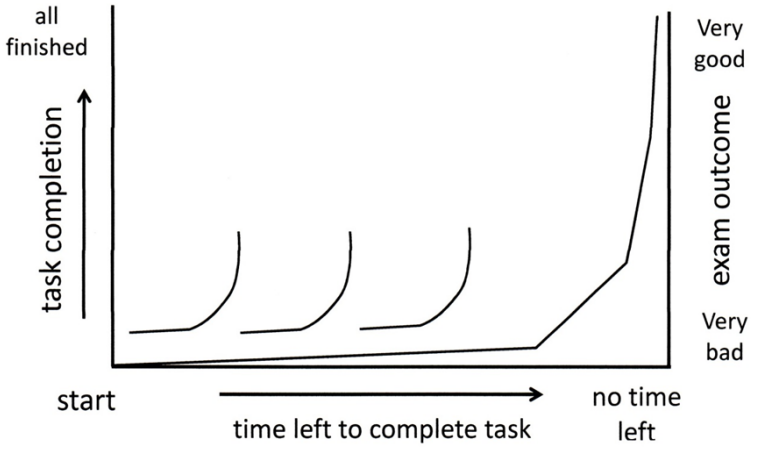

Figure 2: Schematic diagram of the procrastination scallop for students' study behavior with implications for exam outcomes based on inception and completion of academic tasks for one exam plus mini-scallops for weekly quizzes (after Michael, 1993, p. 114).

Assessment

In the past two decade or so, higher education has witnessed a sudden increase of interest in assessment (Dunn, Baker, Mehrotra, Landrum, and McCarthy 2013). Indeed, interest is moving from assessment to assessment outcomes (Naumenko, Hulleman, and Patterson 2013). As there are different types of assessment, assessment of learning outcomes can be achieved with a variety of approaches and tools (Mertler 2003), depending on the objective and level of assessment. At the course level, for example, the objective might be to assess critical thinking in which case short writing exercises might be appropriate, as opposed to assessment of knowledge and understanding, which might require a pretest/posttest approach. Course-level assessment of knowledge and understanding using this approach in a typical college course may take the form of standardized diagnostic tests as is commonly used in Physics (e.g., Coletta, Phillips, and

Steinert 2007; Hake 2001) or nonstandardized, instructor-developed tests (see Mertler 2003). A standard measure in such pre/post assessment is the normalized gain (Bao 2006; Coletta et al. 2007; Hake 2001):

$$g = \frac{\text{Posttest Score} - \text{Pretest Score}}{\text{Maximum Score} - \text{Pretest Score}} \quad (1)$$

I have adopted the pre/post approach for summative assessment in my courses for some years now and have recently examined the data for a systematic and broader impact of my teaching and on my students' learning outcomes, motivated by a scholarship of teaching and learning (SoTL) perspective (Gurung and Landrum 2013). In answering the questions of my teaching effectiveness, my students' learning, and how I could demonstrate these, all things being equal, the pre/post assessment has proven most useful particularly with respect to the use of weekly testing in my courses. Given the dual rationales in behavior and cognition that establish the usefulness of frequent testing for learning and retention, having an established process in place in the form of pre/post assessment provides empirical, as opposed to anecdotal, evidence albeit from a quasi-experimental design. In earlier years I did not implement weekly testing in the introductory course as I had in the more advanced psychology courses, even though I conducted the pre/post assessment in all of them. Demonstrable gains in learning outcomes that may not be in accord with affective student reports in these courses would justify the appropriateness and value of the tests if indeed they made a difference.

Method

Participants

Students enrolled at John Carroll University, Cleveland, OH in three different courses (Introductory Psychology, N = 149; Experimental Design in Psychology, N = 37; and Learning and Behavior, N = 75) completed the assessment instruments in six semesters from the fall of 2009 to the spring of 2012. The experimental design course was offered only in the spring semesters. Our Institutional Review Board (IRB) approved dissemination of the data.

Materials

The assessment instruments consisted of content based 40-item multiple-choice questions randomly selected and screened for accuracy from substitute test banks for the introductory and learning and behavior courses. For the experimental design course, there were 60 items selected as described above. Substitute test banks were used to preclude teaching to the test. Quiz questions were selected from publisher test banks for the assigned textbook for each course during their respective semesters.

Procedures

In each course, the pretest was administered during the first week of classes before any coverage of content. The posttest was administered at the end of the semester, usually during the last week of classes. Students did not know ahead of time when the pretest or the posttest was going to be administered. It was essential that tests were not completed anonymously as the pretest and the posttest needed to be matched for each student (see Hake 2001) to determine $\boldsymbol{g}$. Only matched pairs of pretests and posttests are included in the analysis. Students were advised their performance on the tests would not affect their course grade, but should do their best; they could

earn up to 10 points for each quiz completion and the best ten scores counted toward their course grade. Students thus earned points that cumulated up to 1/6 of their course grade, depending on the course.

From the fall of 2009 to the fall of 2010, students in all sections of the Introductory Psychology course did not complete weekly quizzes as did students in the same course who completed weekly quizzes out of class on the blackboard platform from spring 2010 to spring 2012. Students in the remaining two courses completed weekly quizzes for the entire period of the study. When weekly quizzes were administered for each course, they were worth 10 points each. With usually about 12 such quizzes per semester, only the highest 10 quizzes counted toward course grade.

Results

Average gain in learning was determined for each course using Equation 1. Figure 3 presents the mean gain (g) as a function of the administration of weekly quizzes. For the Introductory Psychology course for which the same assessment instrument was used, mean gain was substantially lower without weekly quizzes (M = 6.09, SD = 18.16) than with weekly quizzes (M = 18.22, SD = 15.97), $t(147) = 3.658$, $p = .0004$, $d = .69$. Mean gains were similarly high for the Experimental Design (M = 25.54, SD = 16.76) and the Learning and Behavior (M = 21.32, SD = 20.71) courses in which weekly quizzes were implemented; these gains are better compared to the no-weekly-quiz gains of the introductory course, $t(146) = 5.747$, $p < .0001$, $d = 1.09$ and $t(192) = 5.440$, $p < .0001$, $d = .79$, respectively.

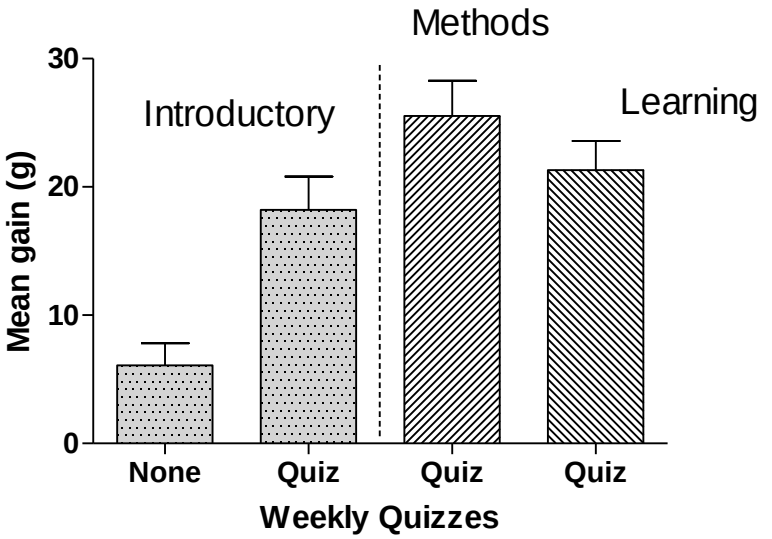

Figure 3: Mean gain (g) as a function of administration of weekly quizzes for the Introductory, Experimental Design (Methods), and Learning and Behavior (Learning) courses. See text for details.

Discussion

When no weekly quizzes were offered, mean gains in learning was quite low compared to when weekly quizzes were administered. This was particularly true in the introductory psychology course in which a direct comparison of availability of weekly testing was possible. The other two courses showed equally high gains in learning comparable or better than those recorded for the introductory course with weekly quizzes. Cohen's effect sizes (Hake 2001) for these comparisons were medium or better. Although no equivalent condition without weekly quizzes was available for the advanced courses, the gains recorded are suggestive and support the general notion that the adverse educational effects of procrastination were mitigated by opportunities for rehearsals afforded by intermediate frequent testing on relevant content.

Not studying as time passes often results from competing academic and/or nonacademic activities, which can be numerous in the typical college student's situation, occupying that time (Michael 1993). Often the instructor has no direct control or influence over those activities and over their intrusion into the relevant study behaviors, safe course-grade relevant requirements. Weekly quizzing serves the purpose of course-relevant activities, filling the procrastination gap with the mini scallops they engender (see Figure 2). An important feature of the value of frequent testing is making their completion and performance count toward the course grade. As noted above, in the present study, students earned points that cumulated up to 1/6 of their course grade. Michael observed that "…if the course grade is of little importance to the student, then the possibility of receiving a low exam score will certainly not function as a basis for aversive control, and studying as a form of escape behavior will certainly not compete with behavior related to other sources of reinforcement" (1993, 115) such as might be available from extracurricular activities.

The weekly quizzes used in the present study, by virtue of being administered out of class and completed individually by the students, meet the definition of practice testing for which there is an abundance of literature on the beneficial impact on learning in terms of frequency and timing (see Dunlosky et al. 2013, for a review); higher frequency and spaced testing appear to be most profitable for learning. This is related directly with distributed practice: According to Dunlosky et al., "In general, distributed practice testing is better than distributed study…" (2013, 37), perhaps due to integrative cognitive processes that occur between practice times and the time of final examination (Bahrick 1979; Cepeda et al. 2006, 2009; Craik 2002; see also Dunlosky at al. 2013, 30).

As interest is moving from assessment to assessment outcomes (Naumenko et al. 2013), taking a SoTL perspective is valuable to the extent that it provides know-how and guidance to interested others, as well as evidence of effectiveness. A limitation of a SoTL approach, however, is the tension between the need to establish empirical evidence and the pragmatic need to demonstrate and/or maintain effectiveness in teaching and learning. For example, in the present case, it might be desirable to collect control data in the experimental design and learning courses to bolster the empirical effect reported for the introductory course, but collecting such data posed a practical challenge of implementing something one knows is potentially detrimental to the particular group of student participating. This is always the difference between basic and applied research: in applied settings, one is not always "free" to demonstrate the same standards of rigor as one could in the laboratory, sometimes for ethical considerations.

Another related factor is the quasi-experimental nature of the present study. Several aspects of the study fit the design. First, although the administration of assessment instruments were planned for each course, students in each course were self-selected into their respective course sections and therefore could not have been randomly assigned to the conditions. Second, the pre/post nature of the assessment present serious questions about threats to internal validity,

including maturation, history, selection, and attrition, all of which could not be controlled for via randomization; attrition was addressed by using only matched pre/post assessment scores and excluding students who did not complete both tests. Naumenko et al.'s (2013) recent report suggests an efficient and effective strategy for eliminating the maturation problem inherent in pre/post assessment by comparing scores from contiguous groups of incoming and outgoing students. Pretest scores of incoming students in a career development course were compared against the posttest scores of outgoing students in a series of back-to-back sessions, revealing the stability of the pretest scores with demonstrably higher posttest scores across sessions. Incidentally, this approach suggests a potential solution to the problem of not having direct comparisons for the experimental design and learning courses in the present study, as the incoming class's pretest would be adequate for comparison with the class that just completed weekly testing the previous semester, without having to run those courses without weekly quizzes, simultaneously addressing any ethical concerns on implementing an "inactive treatment."

Finally, frequent testing may be perceived by students as 'busy work" and therefore undesirable (Dunlosky et al. 2013), but the evidence is overwhelming that it promotes learning and longer-term retention than the alternative of leaving them to their own devices, including last-minute cramming for examinations. For that reason alone, it is well worth it to consider such a strategy for improving learning outcomes. The flip side of the equation is that the faculty may be reticent in adopting such strategy on account of extra preparations and grading. The advantage of practice testing, however, is that they are typically not conducted in the classroom and grading can be completed online when implemented on a platform like blackboard or in conjunction with publishers' supplements on their websites. Engaging students beyond the classroom lecture and activity on an individual basis with some stakes for their final grade ensures they would reap the benefits of all potential cognitive and behavioral processes they encounter as a result. In doing so, the faculty stand to gain from implementing relatively low-cost effective instructional practices that students may benefit from.

REFERENCES

Bahrick, Harry P. 1979. "Maintenance of knowledge: Questions about memory we forgot to ask." *Journal of Experimental Psychology: General* 108: 296-308.

Bao, Lei. 2006. "Theoretical comparisons of average normalized gain calculations." *American Journal of Physics* 74: 917-922.

Cepeda, Nicholas J., Pashler, Harold, Vul, Edward, Wixted, John. T., and Rohrer, Doug. 2006. "Distributed practice in verbal recall tasks: A review and quantitative synthesis." *Psychological Bulletin* 132: 354-380.

Cepeda, Nicholas J., Coburn, Noriko, Rohrer, Doug, Wixted, John T., Mozer, Michael C., and Pashler, Harold. 2009. "Optimizing distributed practice: Theoretical analysis and Practical Implications." *Experimental Psychology* 56: 236-246.

Coletta, Vincent P., Phillips, Jeffrey A., and Steinert, Jeffrey J. 2007. "Why you should measure your students' reasoning ability." *The Physics Teacher* 45: 235-238.

Craik, Fergus I. M. 2002. "Level of processing: Past, present…and future?" *Memory 10*: 305-318.

Craik, Fergus I. M., and Lockhart, Robert. S. 1972. "Levels of processing: A framework for memory research." *Journal of Verbal Learning and Verbal Behavior* 11: 671-684.

Critchfield, Thomas S., Haley, Rebecca, Sabo, Benjamin, Colbert, Jorie, and Macropoulis, Georgette. 2003. "A half century of scalloping in the work habits of the United States Congress." *Journal of Applied Behavior Analysis* 36: 465-486.

Dunn, D. S., Baker, S. C., Mehrotra, C. M., Landrum, R. E., McCarthy, M. A. 2013. *Assessing teaching and learning in psychology: Current and future perspectives*. Belmont, CA: Cengage.

Dunlosky, John, Rawson, Katherine A., Marsh, Elizabeth J., Nathan, Mitchell J., and Willingham, Daniel T. 2013. "Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology." *Psychological Science in the Public Interest* 14: 4-58.

Ferster, Charles B., Skinner, Burrhus F. 1957. *Schedules of reinforcement*. Acton, MA: Copley Publishing Group.

Gurung, R. A. R., and Landrum, R. E. 2013. "Assessment and the scholarship of teaching and learning." In *Assessing teaching and learning in psychology: Current and future perspectives*, edited by D. S. Dunn, S. C. Baker, C. M. Mehrotra, R. E. Landrum, and M. A. McCarthy, 159-171. Belmont, CA: Cengage.

Hake, Richard R. 2001. *Suggestions for administration and reporting pre/post diagnostic tests*. Retrieved from www.physics.indiana.edu/~hake/

Mawhinney, V. T., Bostow, D. E., Laws, D. R., Blumenfeld, G. J., and Hopkins, B. L. 1971. A comparison of students studying-behavior produced by daily, weekly, and three-week testing schedules. *Journal of Applied Behavior Analysis* 4: 257-264.

Mertler, Craig A. 2003. *Classroom assessment: A practical guide for educators*. Los Angeles, CA: Pyrczak Publishing.

Michael, Jack. L. 1993. *Concepts and principles of behavior analysis*. Kalamazoo, MI: ABAI

Moscovitch, Morris, and Craik, Fergus I. M. 1976. "Depth of processing, retrieval cues, and uniqueness of encoding as factors in recall." *Journal of Verbal Learning and Verbal Behavior* 15: 447-458.

Naumenko, Oksana, Hulleman, Christopher S., and Patterson, Heather J. 2013. "Increasing confidence in assessment results: Quasi-experimental approaches." Poster presented at the 25th annual meeting of the Association for Psychological Science, Washington, DC, May 23-26.

Roediger, Henry L. 2013. Applying cognitive psychology to education: Translational educational science. *Psychological Science in the Public Interest* 14: 1-3.

Wanchisen, Barbara A., Tatham, Thomas A., and Mooney, Susan E. 1989. Variable-ratio conditioning history produces high- and low-rate fixed-interval performance in rats. *Journal of the Experimental Analysis of Behavior* 52: 167-179.

Wiener, Harold. 1962. Some effects of response cost upon human operant behavior. *Journal of the Experimental Analysis of Behavior* 5: 201-208.

ABOUT THE AUTHOR

Dr. Abdulrazaq Imam: Associate Professor, Department of Psychology, John Carroll University, University Heights, Ohio, USA.

The International Journal of Assessment and Evaluation is one of ten thematically focused journals in the collection of journals that support The Learner knowledge community—its journals, book series, conference and online community. The journal investigates the dimensions of educational measurement.

As well as papers of a traditional scholarly type, this journal invites presentations of practice—including documentation of measurement practices and exegeses of the effects of those practices.

The International Journal of Assessment and Evaluation is a peer-reviewed scholarly journal.

ISSN: 2327-7920

ISBN 000-0-00000-000-0