COMPUTER BASED TESTING FOR JAPAN STATISTICAL SOCIETY CERTIFICATE: OUTLINE AND PROBLEM EVALUATION

Hirohito Sakurai[1], Atsuhiro Hayashi[2] and Masaaki Taguri[1,3]
[1]National Center for University Entrance Examinations, Japan
[2]Nagoya Institute of Technology, Japan
[3]Chiba University, Japan
sakurai@rd.dnc.ac.jp

The Japan Statistical Society has launched the certification examinations called Japan Statistical Society Certificate (JSSC) in 2011. JSSC evaluates not only knowledge and ability to utilize statistics but also statistical thinking and reasoning. The test sets (forms) of JSSC are prepared for students (from junior high school to university) and working adults who are interested in and/or need statistics. Among them, Grade 2 and 3 examinations are conducted by paper-and-pencil and Computer Based Testing (CBT) styles with multiple choice. In this paper, we briefly outline the CBTs of JSSC and discuss evaluation methods of testlets and items included in the test forms. These methods are universally applicable and useful to many kinds of test evaluation. Future works relating to the CBTs are also discussed.

INTRODUCTION

Japan Statistical Society Certificate (JSSC) is one of the certification examinations conducted in Japan since 2011, and is managed by the JSSC Examination Center of Japanese Association for Promoting Quality Assurance. It evaluates not only knowledge and ability to utilize statistics but also statistical thinking and reasoning for students (from junior high school to university) and working adults. The test sets (forms) of JSSC are currently composed of Grade 1 (Mathematical Statistics), Grade 1 (Applied Statistics), Grade Pre-1, Grade 2–4, Survey Statistician and Professional Survey Statistician. Computer Based Testing (CBT) is also introduced to Grade 2 and 3 since August 2016. The main target of each Grade is summarized in Table 1. The establishment and mission of JSSC are explained in detail in Yoshizoe (2011, 2013) and Yoshizoe et al. (2013). Detailed explanation of Grade 2 and 3 examinations can be found, for example, in the articles by Imaizumi & Tamura (2011), Takeuchi & Watanabe (2011), Fujii et al. (2014) and Taguri (2016).

Among the above kinds of test forms, Grade 2 and 3 examinations adopt both paper-and-pencil and CBT styles, and the others only paper-and-pencil style. Applicants to these grades can take one or more tests according to one's preference. In this paper, we focus on the CBTs of Grade 2 and 3, and outline the CBTs and evaluation methods of testlets and test items in the test forms. Future works relating to the CBTs are also discussed.

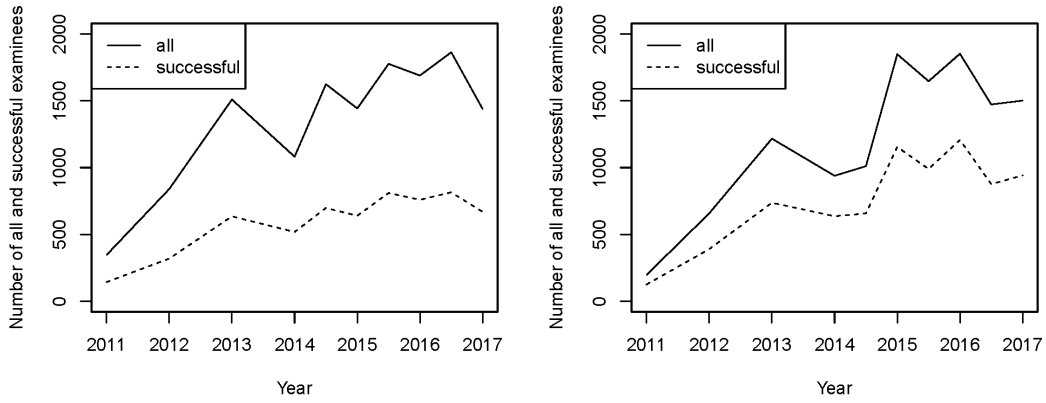
Table 1. Main target of each Grade

Grade	Starting year and month	Main target
Grade 1	2012 November	Undergraduates majoring in subjects where statistical methods are widely used (Higher level)
Grade Pre-1	2015 June	Undergraduates majoring in subjects where statistical methods are widely used
Grade 2	2011 November	University students majoring in sciences
Grade 3	2011 November	Senior high school students and university first graders
Grade 4	2011 November	Junior and senior high school students
Survey Statistician	2011 November	Current or prospective survey enumerators
Professional Survey Statistician	2011 November	Current or prospective managers of statistical surveys

In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, July, 2018), Kyoto, Japan.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org

OUTLINE OF GRADE 2 AND 3 EXAMINATIONS OF JSSC

The full mark of Grade 2 and 3 examinations is 100, and the minimum pass marks of them are confidential in the paper-and-pencil tests; however they are announced in the CBTs: 60 for Grade 2 and 70 for Grade 3, respectively. It is possible for applicants of these grades to take one or both tests according to one's preference. Both examinees are allowed to use a calculator which can calculate arithmetic operation, percentage, and square root. Figure 1 shows the numbers of all and successful examinees of Grade 2 and 3 in 2011–2017. The pass rate of Grade 2 and 3 in 2011–2017 are summarized in Figure 2.



(a) Grade 2 (b) Grade 3

Figure 1. Number of all and successful examinees of Grade 2 and 3 in 2011–2017

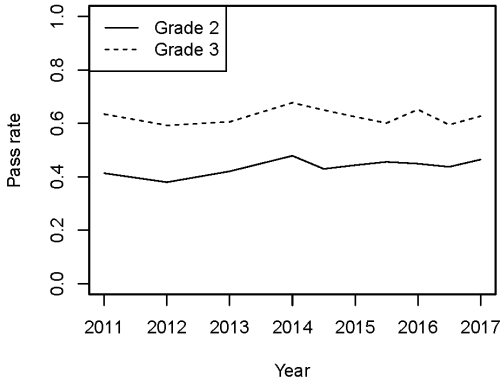


Figure 2. Pass rate of Grade 2 and 3 in 2011–2017

Grade 2 Examination of JSSC

The Grade 2 examination evaluates and certifies the following abilities which are required to master in the first and second years' basic course in universities:

- to find problems to be solved in miscellaneous situations,
- to construct and verify hypotheses for the problems from statistical view points, and
- to solve the problems to find opportunities to acquire new findings.

The test form includes about 30–40 items (problems), and examinees of Grade 2 have to solve them in 90 minutes. As for the answer format of Grade 2 examination, there are five choices of answer candidates for each item, and we choose one correct answer from them. The coverage of Grade 2 (Japan Statistical Society Certificate, 2017a; Yoshizoe et al., 2013) is summarized as follows:

- Collection and compilation of data
- Analysis and presentation of data
  - Descriptive techniques
  - Introductory probability theory
  - Random variables, expectation, distribution of some statistics

- Scatter plot, correlation, simple regression
- Statistical inference: interval estimation
- Statistical inference: testing hypothesis
- Statistical inference: regression analysis, analysis of variance

Grade 3 Examination of JSSC

The Grade 3 examination evaluates and certifies the following abilities of statistical utility required as basic knowledge of statistics in senior high schools:

- statistical literacy: definitions of basic terms and concepts
- statistical reasoning: basic interpretation of terms; relevance of two or more terms or concepts
- statistical thinking: utilization of statistics based on specific context

The test form has about 30 items, and examinees of Grade 3 have to solve them in 60 minutes. Each item has four or five choices of answer candidates, and we choose one correct answer from them. The coverage of Grade 3 (Japan Statistical Society Certificate, 2017b; Yoshizoe et al., 2013) is summarized as follows:

- Basic concepts of statistics (Statistical literacy)
- Interpretation and relationship (Statistical inference)
- Utilization of statistical observations (Statistical thinking)

COMPUTER BASED TESTING IN JSSC

JSSC has introduced Computer Based Testing (CBT) to Grade 2 and 3 examinations since 29 August 2016; however the paper-and-pencil tests are also continued as before. The coverage of the CBT is the same as the corresponding paper-and-pencil test, as described above. Test items in a test set are randomly selected from an item pool on computer; the item pool contains much more items than any single examinee encounters. A test set is different for individuals every time when an examinee takes a test. Examinees are promised to keep confidential about the content of the CBT, because they will know some of the item pool when they take the test.

Development of test forms (1): Grade 2 examination

As for Grade 2 examination, problems (items) included in the test forms are decided based on the 6 sets of examinations (conducted from November 2011 to June 2015) with the following policies:

- The CBT of Grade 2 is subdivided to the following three areas:
  - (A2) Descriptive statistics: one- and two-dimensional descriptive statistics
  - (B2) Probability related: collection of data, probability, probability distribution, and sampling distribution
  - (C2) Inferential statistics: estimation, test (hypothesis test, chi-squared test), and linear models (regression, experimental design)

  A test set is generated by the combinations of areas (A2), (B2) and (C2). As a result, about 35 problems (items) are totally included in a test set of CBT.
- Number of items in each area: approximately the mean of the above 6 sets of paper-and-pencil type examinations
- Difficulty level of items: each area is formed by easy and difficult items in half and half, where difficulty level is judged by reference to the past correct answer rates of the same kind of items in the above 6 sets of examinations
- Number of test forms needed for the operation of CBT (Taguri, 2016; Diaconis and Mosteller, 1989): let m be the number of examinees in an examination room, P be the probability of taking different test forms at all seats (m) in the examination room, and n be the number of test forms needed for the administration, then $\log P \approx -m^2/(2n)$ holds; so we need to prepare $n \approx -m^2/(2\log P)$ test forms in advance where m/n is about 0.15 or less. The specific values of n are given in Table 2. From Table 2, we suffice to prepare 500–1000 test forms. Note that, if there are examinees at both sides of nearest neighbors ($m = 2$), we obtain $P = 0.996$ for $n = 500$ and $P = 0.998$ for $n = 1000$.

Table 2. Number of test forms needed for the operation

$m \setminus P$	0.99	0.9	0.8	0.7
20	19900	1898	896	561
15	11194	1068	504	315
10	4975	475	224	140

Development of test forms (2): Grade 3 examination

Almost the same policies as Grade 2 examination are applied in development of Grade 3 examination. The only difference is that the CBT of Grade 3 is subdivided to the following three areas:

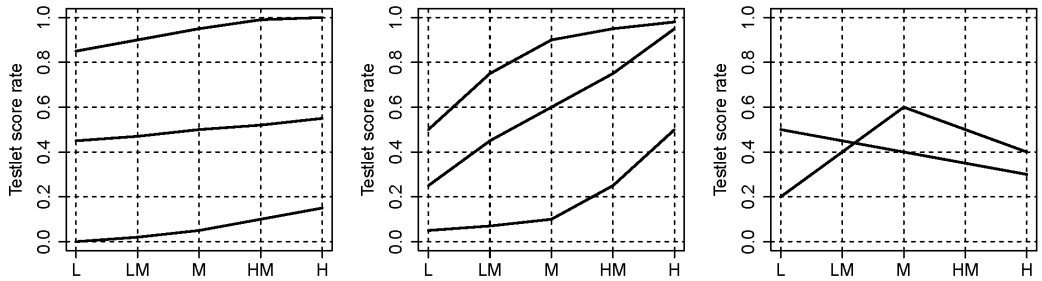
- (A3) Types of data, sample survey, experimental survey, and statistical graphs
- (B3) Data aggregation, representative values of data, and variation of data
- (C3) Probability and time series

A test set is generated by the combination of areas (A3), (B3) and (C3). As a result, examinees have to solve totally about 30 problems (items) in the CBT of Grade 3.

EVALUATION OF TEST ITEMS IN CBT OF JSSC

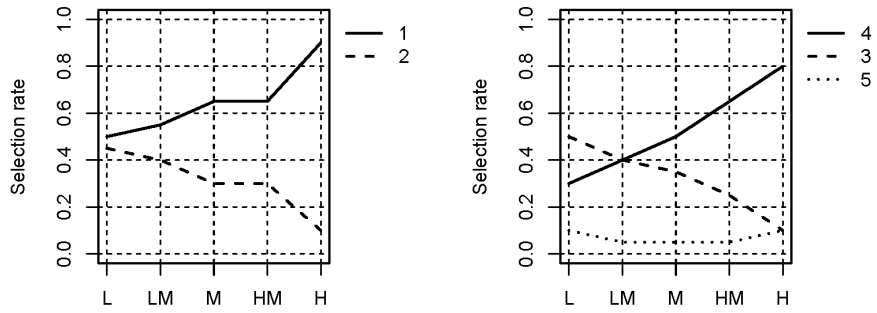
For the CBTs of Grade 2 and 3, we cannot decide pass/fail judgment considering the difficulty level of the examination items because the minimum pass marks of them have been announced. We, therefore, constantly need to check the pass criteria, and to modify and/or replace the test problems (items) included in the item pool as appropriate. To be more precise, we must regularly monitor correct answer rates of every item used in the CBTs. If there are items with extremely high/low correct answer rate, we need to investigate the causes, and to modify or replace the items in some cases. We also have to pay attention to the difficulty level of combination of items as a set of an examination. If necessary, we have to modify and/or replace the combinations of items in (A2), (B2), (C2) or (A3), (B3), (C3), respectively. Moreover, once in several years, we need to add and/or replace quite a few items since it is desirable to have items in the item pool as timely as possible. In the following part, we will give some analytic techniques which are kinds of like empirical versions of problem evaluation by item-response curve. They are applicable to wide-ranging studies and may be effectively used in miscellaneous test evaluations.

In order to evaluate behavior of testlets and items included in a test set of CBT, we apply statistical analysis technique based on classical test theory, where a testlet is a set or group of test items classified as the same area. Our analyses of the CBTs stated below use testlet scores and 0/1 scores of items, where testlet score is a weighted sum of scores on several items belonging to the same area, and scores 1 and 0 represent correct and incorrect response of examinees to each test item, respectively. Results of data analysis of Grade 2 and 3 examinations with paper-and-pencil style are given, for example, by Iwasaki (2012), Fukasawa (2012) and Tarumi (2013).



(a) Low discrimination (b) High discrimination (c) Need to consider

Figure 3. Examples of quintile testlet response chart



(a) Almost two choices (b) Almost three choices

Figure 4. Examples of quintile item response chart

Testlet analysis

For evaluation of testlets, mean, testlet score rate, standard deviation, maximum and minimum are calculated for each area based on the testlet score. We also make a parallel box-and-whisker plot for each area, and carry out correlation analysis among the areas (A2)–(C2) or (A3)–(C3), that is, we calculate correlation coefficients and draw scatterplot matrices, among the areas. Further, as an evaluation method of a testlet, a line graph called *Quintile Testlet Response (QTR) chart* is used. Typical examples are given in Figure 3. The horizontal axis represents groups of examinees divided into 5 groups by the quintiles of total score of examinees; the groups are expressed as L (low), LM (low-middle), M (middle), HM (high-middle) and H (high), respectively. The vertical axis represents testlet score rate in each group. If a testlet distinctively discriminate all groups, the line graph monotonically increases from L to H and the corresponding QTR chart will show the pattern as the middle line in Figure 3 (b); the upper and lower lines in Figure 3 (b) distinctively discriminate lower (from L to M) and higher groups (from M to H), respectively. The patterns in Figure 3 (a) are examples of low discrimination; the upper, middle and lower lines are corresponding to easy, intermediate and difficult items for examinees. If the patterns such as Figure 3 (c) appear, we should pay attention to the corresponding items because we need to consider the reasons why such strange response patterns are obtained.

Test item analysis

Item difficulty and item discrimination are indicators for evaluation of test items. The former is defined by correct answer rate of an item, and the latter by item-total correlation which is a correlation coefficient between the response 0/1 of an item and total number of correct answers in a test form. Further, a line graph called *Quintile Item Response (QIR) chart* is used for item evaluation (Tarumi, 2013). This is drawn by the same idea as the QTR chart. The horizontal and vertical axes are groups from L to H, and selection rate of correct or incorrect answer for each group, respectively. Some examples are shown in Figure 4. It is possible to draw all lines corresponding to multiple choice categories in a QIR chart, however, to make it easy to see, it is usual to only draw the lines corresponding to a correct answer and incorrect answers selected by 10% or more examinees. Figure 4 shows almost two or three choices examples though there are five choices (1–5) in each item, because their sum of selection rate in each group is about 0.9–1; the correct and incorrect answers are expressed by the solid and other lines, respectively.

FURTHER DEVELOPMENTS

For further developments, improvement plans of JSSC are now being discussed among stakeholders. The coverage of Grade 3 and 4 is designed to correspond to Official Curriculum Guidelines (OCGs) for senior high school and junior high school, respectively. The OCGs are controlled by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) in Japan, and the MEXT emphasizes the importance of statistical thinking and data analysis, and has plans to revise OCGs for elementary, junior and senior high schools. Although the applicants of JSSC are not restricted as such students, the revision of the coverage of Grade 2–4 will be discussed taking account of the revisions of OCGs as a good opportunity. Based on new coverage

of JSSC and on evaluation results of CBTs, the items of Grade 2 and 3 CBTs will be added and/or modified. Further, JSSC has a plan to introduce CBT for Grade 4 and Survey Statistician. Since the CBT system makes it possible to set up test dates and test sites more flexible than paper-and-pencil test, it is preferable for children and students at the primary and secondary education stage to take CBT with less time and place constraints.

REFERENCES

Diaconis, P. and Mosteller, F. (1989). Methods for Studying Coincidences. *Journal of the American Statistical Association*, *84*, 853–861.

Fujii, Y., Fukasawa, H., Takeuchi, A. & Watanabe, M. (2014). A Certification System for Statistics Knowledge and Skills by Japanese Statistical Society. *iase-web.org/icots/9/proceedings/pdfs/ICOTS9_3B1_FUJII.pdf*.

Fukasawa, H. (2012). Analysis of Grade 3 and 4 Examinations' Results of Japan Statistical Society Certificate. *Estrela*, *219*, 12–17. (in Japanese)

Imaizumi, T. & Tamura, Y. (2011). On Grade 2 Examination of Japan Statistical Society Certificate. *Estrela*, *210*, 11–15. (in Japanese)

Iwasaki, M. (2012). Analysis of Grade 2 Examination Results of Japan Statistical Society Certificate. *Estrela*, *219*, 6–11. (in Japanese)

Japan Statistical Society Certificate (2017a). Coverage of Grade 2 Examination of Japan Statistical Society Certificate. *www.toukei-kentei.jp/wp-content/uploads/grade2_hani_170727.pdf.* (in Japanese)

Japan Statistical Society Certificate (2017b). Coverage of Grade 3 Examination of Japan Statistical Society Certificate. *www.toukei-kentei.jp/wp-content/uploads/grade3_hani_170727.pdf*. (in Japanese)

Taguri, M. (2016). Introduction of Computer Based Testing to Japan Statistical Society Certificate. *Estrela*, *270*, 8–13. (in Japanese)

Takeuchi, A. & Watanabe, M. (2011). Goal and Contents of Japan Statistical Society Certificate Grade 3 and 4 Examinations: Statistical Thinking as a Minimum Standard Skill for Civic Society. *Estrela*, *210*, 16–22. (in Japanese)

Tarumi, T. (2013). Analysis of Grade 2 Examination Results of Japan Statistical Society Certificate in 2011: From a Perspective of Quintile Item Response Chart. In Japanese Inter-university Network for Statistical Education (ed.), *Report on Quality Assurance of Statistical Education and Analysis of Examination Problem.* Tokyo: Japanese Inter-university Network for Statistical Education, Chapter 6, 77–92. (in Japanese)

Yoshizoe, Y. (2011). Establishment and Mission of Japan Statistical Society Certificate. *Estrela*, *210*, 2–10. (in Japanese)

Yoshizoe, Y. (2013). Establishment and Mission of Japan Statistical Society Certificate. In Japanese Inter-university Network for Statistical Education (ed.), *Report on Quality Assurance of Statistical Education and Analysis of Examination Problem.* Tokyo: Japanese Inter-university Network for Statistical Education, Chapter 1, 3–11. (in Japanese)

Yoshizoe, Y., Takemura, A. and Kawasaki, S. (2013). Quality Assessment of Statistical Education by Japan Statistical Society. Presented at IASE/IAOS Joint Satellite Conference, 22–24 August 2013, Macau SAR. *iase-web.org/documents/papers/sat2013/IASE_IAOS_2013_Paper_1.4.2_Kawasaki_ppt.pdf.*