Analysis of a Test Items Instrument using CTT and IRT upon the Competence of Vocational School Students in Digital Electronics

N. Kholis[a], B. Suprianto[b], Munoto[c], [a,b,c] Electrical Engineering, Departement Of Engineering, Universitas Negeri Surabaya, Email: nurkholis@unesa.ac.id

This research has the objective to analyse and synthesise instruments on the test items by using the Classical Theory of Tests (CTT), and Item Response Theory (IRT) to determine the competence of students in digital electronics in the vocational school setting. The methods used in conducting the research, make the item in order to examine the test items on the competence of Digital Electronics. The test consists of 40 items in a multiple choice format (var1-var40), and includes the following topics: (1) interpret models of atoms of semiconductor material, the characteristics of the diode, and applying the diode; (2) test the special diode, such as the diode LED, the transistor in the electronic circuit, and determine the working point on the transistor; (3) applying the Boolean algebra on the gates of digital logic, and kinds of logic gate, and (4) build a variety of gates in the basic logic circuit. The respondents were comprised of 704 students. Using the CTT to produce the Kuder-Richarson coefficient of reliability/KR-20 (0.6281), number of items in the scale 40, the number of complete observations 704 and benchmark items that can be used is the value above 0.2 with a look at the column item-rest correlation as much as 4 items, includes: var4 (0.2480); var35 (0.2128); var36 (0.2095); and var37 (0.2170). And items that can't be used under 0.2 for theory classic test contained 36 items. Using IRT based on the observations on the unweight fit (Outfit mean square) has a value above 0.5 and the weight fit (infit mean square) value is below 1.5, the grains that can be used contained 40 items.

Keywords: *Theory of classic test, Item response theory, Competence, Outfit mean square, Infit mean square.*

Introduction

The Classical Test Theory (CTT) has been widely used in the process undertaken upon an analysis item. In determining the popularity, it can be used because of the possibility of excess and deficiency, which is obtained in the CTT. In excess and deficiency, it can be assumed at the level of difficulty, and the power of discrimination upon the items in a CTT, which can be used by calculating manually. In the counting, it can be obtained manually, as in the analysis using the CTT based on the data obtained with the amount that is not too much. In addition, that the excess of the CTT is not in spite of our weaknesses, and shortcomings. For example, the weaknesses, and shortcomings at the level of difficulty, and power of discrimination against the items obtained depend on the sample (Hambleton & Swaminathan, 1985). The dependence of the sample used to cause the characteristics on the item, which is analysed with CTT, can be changed according to the context of the respondents. The meaning of such a term refers that an item can have a low level of difficulty because the item is completed by the group of respondents with a high ability. Thus, the level of difficulties on these items can be high, when completed by the group of respondents with a low ability. The weakness, and other deficiencies of the CTT, is that it is more oriented on the test compared to an item (Hambleton, Swaminathan, & Rogers, 1991). Therefore, the researchers seek to investigate the two theories of the CTT, and the Item Reponse Theory (IRT) upon the competence of vocational school students in digital electronics.

This study examines the following objectives:

1. Analyse and synthesise test items using the CTT upon the competence of vocation school students in digital electronics.
2. Analyse and synthesise test items using the IRT upon the competence of vocational school students in digital electronics.

Literature review

Instruments

The definition of a ‘research instrument’, is a tool used to collect research data, whether the data is qualitative or quantitative. Qualitative data can be in the form of images, words, and/or other objects that are non-numbers. Meanwhile, quantitative data is data that is in the form of numbers. In qualitative research, the instrument is the main researcher. Thus, what is meant by research instruments in this instance, is the instrument of quantitative research.

The quantitative data itself can be grouped into two, namely: the nominal data, and the continuum data. The data is said to be on the level of nominal or a nominal scale if the figure serves to identify, which distinguishes the types of other subjects. The difference in the figures only indicates the presence of the object or the subject of a separate and not equal. Meanwhile, the data continuum consists of large-scale data in ordinal, interval, and ratio forms.

Quantitative data is usually obtained through measurement, i.e. a process of providing figures on the subject, the object or related to the other. Therefore, the research instrument can also be referred to as a measuring instrument, and this instrument can be in the form of tests and nontes. Measuring said test if loading some of the questions to which the answer is no right and wrong. On the contrary, if in the measurement tool of the answer there is no right or wrong choice, it can be referred to as a scale, questionnaire or inventory.

The main requirements of a good instrument is that it is valid, and reliable. The validity of a measuring instrument is the extent to which the instrument was able to measure what should be measured. Validity, in general, is whether the level is or is not present. The validity of an instrument is also only seen from a specific purpose. For example, maknaanya is an instrument that is said to be valid for measuring the attribute ‘X’, but is not valid for measuring the attribute ‘Z’.

According to Allen and Yen (1979), there are three methods commonly used to estimate the coefficient of reliablitas, namely: the re-test method, the parallel test method, and the internal consistency method. In general, each of the three methods will produce estimates of the coefficient of reliability (rx), which are different. Thus, what is produced is only an estimate, as the actual value of this coefficient is not observed.

As the name implies, the re-test method involves taking the same test twice. The results are then correlated and obtain the estimates of reliability. The method of the re-test produces estimates of the reliability of the test, including whether it is reasonable, but this method can have several drawbacks. The first method is potentially affected by the carry-over effect between the tests; the first test is very likely to affect the results of the second test. The second drawback concerns the time of the test.

According to Allen and Yen (1979), there are three commonly used ways to divide the test into two parts. Firstly, the method of odd-evens, grain-grain tests allocates grouping based on the odd numbered items in one group, and the even numbered into the second group. Secondly, the method of splitting the two in accordance with the sequence numbers.

According to the Team Pusisjian (1997/1998), there are six steps to develop measuring instruments: (1) draw up the specification of the measuring instrument, including a grating and indicators; (2) written questions; (3) study questions; (4) test; (5) analyse the grain of the instrument; and (6) assemble the instrument and label the specifications.

The specifications of this instrument include the purpose of the measurement, the grating of the instrument, scale of measurement, and the length of the instrument. Therefore, in determining the specifications of the measuring, it requires determining the purpose of the instrument, developing a grating instrument, specifying a scale of measurement, and determining the length of the instrument. In the context of the future, it has been argued that there are two kinds of instruments: instruments for testing, and nontes. Therefore, there is a need to distinguish between the gratings of the instrument for the test, and the grating of the instrument nontes. In detail, the preparation of both gratings includes grating tests or instruments, and grating non-test instruments.

In a grating instrument or test, after the test purpose is specified, then the next activity is to construct the lattice test. This grating is essentially a table matrix that contains specifications about what will be written. The lattice contains the purpose, competence standard, basic competence, subject matter, and the assessment, which holds the form, and the type of bill. The standard of competence is translated into a basic competence. Basic competence is broken down into several indicators, and of the indicators, a grain of the instrument is made. There are three steps that must be met to write the grating: choose a standard of basic competence, write the indicators, and determine the form of the test. Broadly speaking, there are two forms of the test that is widely used by educators. Namely, the shape of the objective, and the form of description or non-objective. Several of the test items have the possibility that each form of the test has its advantages, and disadvantages.

In regard to grating non-test instruments, the preparation begins with the determination of the conceptual definition, which is then translated again to the operational definition. From the operational definition, it is then translated into several indicators, which are further elaborated into a grain of the instrument. As has been described previously, the instruments of the non-test are divided into two: the scale, questionnaires, and inventories.

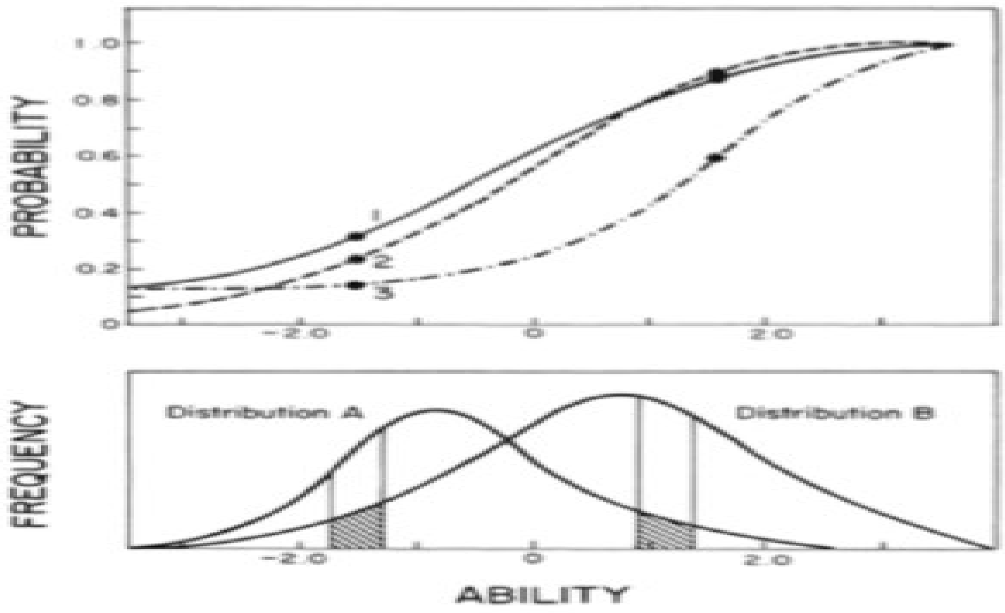
Classical Test Theory (CTT)

The CTT does not pay attention to how the respondents respond to the items. As for how to apply on the ability of the respondents viewed in accordance with the total score of the number of correct answers of the respondents, i.e. without distinguishing whether the items are answered correctly by the respondents is the item easy or difficult. The CTT has a score model true otherwise $X = T + E$, with X as a variable is a manifest or also called scores it is observed that obtained from the grain test, T as a latent variable also called score true that not seem, and E as a component of the errors (Croker & Algina, 1986). In the process of selection on the grain of the 'ng', it is based on the correlation between grains suggest and defend against some of the grain has high correlation with another item and throw a few grains that are low correlated. The weaknesses, and shortcomings of the CTT drive the emergence of the IRT.

Item Response Theory (IRT)

The IRT has quickly become the main flow of the test psokologi, and education. The IRT is a measurement-based model with estimates of the level of ability brgantung on the response of the test participants, and the parameters of the grains of the test. It is compiled from the understanding that the probability of the respondents answering correctly to an item, which can be described as a simple function of the position of respondents on a trait latent, coupled with one or more parameters, has become the characteristics of the items (Molenaar, 1995). Hambleton, Swaminathan, and Rogers (1991) stated that the basis of the IRT is firstly, the performance of the respondents against the item that can be predicted, based on a number of factors called trait or abilitas latent, which shows the capabilities or characteristics of the properties. Secondly, the relationship of the performance of the respondents to the item, and the trait underlying the performance on the items described, increased in a monotonic function, a form called the Item Characteristic Curve (ICC). The ICC of the three items is shown in Figure 1. From the charts, it can be known that the probability of a respondent answering an item correctly depends on the abilitas, regardless of whether the respondents are derived from group A or B.

Figure 1: The probability of answering three of the items in group A, and B



(Hambleton & Swaminathan, 1985)

The Competence of Digital Electronics

Competence is an ability to perform a job or task based on the skills, knowledge, and attitudes required of a particular job. The knowledge required and related to the work includes know and understand the knowledge of each field; and determine the knowledge associated with the rules, procedures, and new techniques in government institutions. Reagdring the skills required of the individual, they include the ability to communicate well in writing, and the ability to communicate clearly, orally. As for the attitude of the individual, it should include an ability to communicate in creativity in the work, and the existence of high morale.

In terms of the competency in digital electronics which is used in this research, there are four basic competencies. Firstly, a collaboration model of the atom semiconductor material, the characteristics of the diodes, and applying a diode. This may encompass: a set of semiconductor P-Type and N-Type; build the rectifier circuit half wave rectifier single phase; building a circuit of full wave rectifier single phase; connect the prisoner in the zener diode with the output voltage of the load; designing the circuit of the voltage stabiliser in parallel using the zener diode; planning a zener diode for the purposes of the reference voltage; analyse the datasheet of zener diodes to determine custody in, and dimensional stability of the circuit; experiment circuit of the voltage stabiliser using zener diode; and interpret the measurement data. Secondly, test the special diode, such as the diode LED, the transistor in the electronic circuit, and determine the working point on the transistor. This may encompass: conducting special diode experiments, such as LED diodes, varaktor, Schottky, PIN, and tunnel interpretation of data measurement results; experiment with a bipolar transistor, as an amplifier single one the signal level using the software; perform the bipolar transistor experiment as a device switch by using software; analyse the placement of the working point DC transistor; analysing the technique of the bias voltage divider of the transistor circuit; analyse the techniques of the bias current feedback, and voltage series transistors. Thirdly, applying the Boolean algebra on the gates of digital logic, the kinds of logic gate, and apply it in context. This may encompass: combining two elements of a binary on the system, which is the summation of Boolean algebra; combining a series of gates of digital logic with Boolean algebra; build the basic principles of some of the logic gate AND – OR – NOT – NAND – NOR – XOR – XNOR; build the basic principle of a logic gate exclusively as OR-NOR; and analyse the timing of the output of some of the logic gate. Fourthly, build a variety of gates on the basic logic circuit. This may encompass: experiment several gates of basic logic, AND – OR – NOT – NAND – NOR, using the software and measurement hardware, as well as the interpretation of data measurement results; perform the experiment logic as exclusively OR – NOR, using the software and measurement hardware, as well as the interpretation of data measurement results; and analysing the timing of the output of some of the logic gate.

The number of indicators in each competency policy is, as follows: KD 1: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, and 14; KD 2: 15, 16, 17, 18, 19, and 20; KD 3: 21, 22, 23, 24, 25, 26, 27, 28, 29, and 30; and KD 4: 31, 32, 34, 35, 35, 36, 37, 38, 39, and 40.

A few number of indicators such as an instrument that consists of several items. The Item has a number of 40 items to be tested the level of difficulties, so we get items that can be used as instruments on the ability of the theory tests the student vocational.

Methodology

In conducting this study, the researchers create the item, in order to examine the test items against a competency in digital electronics. Such tests consist of 40 items of multiple choice (var1-var40). Test the access to some of the topics include: (1) interpret models of atoms of semiconductor material, the characteristics of the diode, and applying the diode; (2) test the special diode, such as the diode LED, and the transistor in the electronic circuit, and determine the working point on the transistor; (3) apply the Boolean algebra on the gates of digital logic, and kinds of logic gate; and (4) build a variety of gates on the basic logic circuit. The number of respondents was 704 students. A test analysis was conducted to estimate the parameters of the hardship clause (p), which was based on the CTT. Meanwhile, the ConQuest was used to estimate the parameters of the hardship clause (b), which was based on the IRT. This was undertaken to analyse the instruments of competency in digital electronics. The results of the test were analysed by comparing both theories — CTT, and IRT. The item test used in this study was in accordance with the model of Rasch. A grain test made selections based upon the value of the IMS, and OMS. Linacre (2002) compiled a table to interpret the meaning of the value of the IMS, and OMS. On the following table shows the meaning of the values of the IMS, and OMS. The values of IMS, and OMS range from 0.5 to 1.5.

Table 1: The range of values on the IMS and OMS

Value	**Implications for the measurement**
>2.0	Damage measurement system
1.5–2.0	No useful meaning for the measurement
0.5–1.5	Beneficial for measurement
<0.5	Not useful for measurement but it does not spoil

Results and Findings

The variables used consist of var1 until var40. The total number of responses was 704. The testing was first carried out by using an item analysis in the form of the CTT.

In analysing the use of the CTT upon competency in digital electronics, the Kuder-Richarson coefficient of reliability/KR-20 (0.6281), number of items in the scale 40, the number of complete observations 704 and benchmark items that can be used is the value above 0.2 with a look at the column item-rest correlation as much as 4 items, includes: var4 (0.2480); var35 (0.2128); var36 (0.2095); and var37 (0.2170). Reviewing the Table 1, there were 36 items which cannot be used in the CTT. These items comprise: var2 (0.1511), var3 (0.1950), var5 (0.1996), var6 (0.1796), var7 (0.1611), var8 (0.1427), var9 (0.1061), var10 (0.1088), var11 (0.1268), var12 (0.1362), var13 (0.1174), var14 (0.1644), var15 (0.1395), var16 (0.1323), var17 (0.1286), var18 (0.1053), var19 (0.1225), var20 (0.1405), var21 (0.1572); var22 (0.1797), var23 (0.1537), var24 (0.1210), var25 (0.1122), var26 (0.1180), var27 (0.1239), var28 (0.1420), var29 (0.1822), var30 (0.1531), var31 (0.11), var32 (0.1868), var33 (0.1903), var34 (0.1858), var38 (0.1627), var39 (0.1525), and var40 (0.2460). Meanwhile, the level of difficulty of the items can be seen in the table of item difficulty. The reliability test produced values of 0.6273, with the number of items in the scale 40, and the inter-item correlation average of 0.0404.

In analysing the use of IRT upon the competency in digital electronics on the multiple choice questions, the results of the item analysis multiple-choice response model parameter estimates pay attention to the column range of values unweighted fit, and weighted fit (Linacre).

If in analysing the use of the rasch analysis with the provisions of 0.5–1.5, it is beneficial for measurement. Based on the results of the item analysis using the rasch software Conquest, it can be concluded that the value of unweighted fit (outfit mean square), and weighted fit (infit mean square) on all of the items can be used. The value of infit mean squared (IMS) on all items has the value of the minimum of 0.97, and a maximum value of 1.02. Meanwhile, the value of the outfit mean squared (OMS) on all items has the value of the minimum of 0.97, and a maximum value of 1.02.

In analysing the use of the rasch analysis to have a provision of 0.5–1.5 with the implications of the measurement of the ‘benefit measurement’. Based on the results of the item analysis using the rasch software Conquest, it can be concluded that there are some items which cannot be used in the measurement of a number of zero. As for the analysis items which can be used, the clarifications obtained are as follows: (1) based on the observations in the table, that when the unweight fit has a value above 0.5, and the weight fit value is below 1.5, then all the grains can be used; and (2) it can be concluded that all the grains can be used.

Conclusion

This research has resulted in two conclusions, as follows:

1. use the CTT/ Clasical Test Teory to produce, Kuder-Richarson coefficient of reliability/KR-20 (0.6281), number of items in the scale 40, number of complete observations 704 and the size of the item that can be used is the value above 0.2 with a look at the column item-rest correlation as much as 4 items, which can not be used in the theory of classical test as many as 36 items, the reliability test produces a value 0.6273, inter-item correlation average 0.0404.
2. Using the IRT, there are 40 items which are unweight fit at above 0.5, and there were zero items in the weight fit value, which is below 1.5. Thus, all 40 items can be used.

REFERENCES

Allen, M.J. & Yen, W.M. (1979). Introduction to measurement theory. Monterey, CA: Brooks/Cole Publishing Company.

Everitt, B., & Howell, D. C. (Eds.). (2005). Encyclopedia of statistics in behavioral science. Hoboken, N.J: John Wiley & Sons.

Furr, R. M. (2011). Scale construction and psychometrics for social and personality psychology. London: SAGE.

Hambleton, R. K., & Swaminathan, H. (1985). Item response theory, principles and applications. New York: Springer Science+Business Media.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. California: Sage Publications, Inc.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. New Jersey: Lawrence Erlbaum Associates.

Molenaar, I. W. (1995). Some background for item response theory and the rasch model. Dalam G.H. Fischer, & I. W. Molenaar (Eds). Rasch models. New York: Springer-Verlag.

Nunnally, J.C. (1978). Psychometric theory. New York: McGraw Hill Book Company. Inc.

Ridho, A. (2007). The characteristics of the psychometric test approach based on test theory classical and theory of the response item. Journal Of Psychology INSAN, 2(2), 1-27.

Snipes, M., & Taylor, D. C. (2014). Model selection and Akaike Information Criteria: An example from wine ratings and prices. Wine Economics and Policy, 3(1), 3–9. https://doi.org/10.1016/j.wep. 2014.03.001

Sumintono, B, & Widhiarso, W. (2013). The application model of rasch for social sciences research. Trim Komunikata Publishing House.

Swaminathan, H., Hambleton, R. K, & Rogers, H. J. (2007). Assessing the fit of item response models. Handbook of Statistics, 26, 683-718.

Veerkamp, W. J. & Berger, M. P. (1999). Optimal item discrimination and maximum information for logistic IRT models. Applied Psychological Measurement, 23(1), 31-40.

Wiberg, M. (2004). Classical test theory vs. item response theory. Umea, 10(5), 1–27.