Development of Learning Outcomes Assessment Instruments for Fashion Technology Courses

Wulansari Prasetyaningtyas[1*], Widihastuti Widihastuti [2], Edi Istiyono[2]

[1] Fashion Education Department, Faculty Engineering, Universitas Negeri Semarang, Semarang, Indonesia
[2] Educational Research and Evaluation, Universitas Negeri Yogyakarta, Yogyakarta, Indonesia
Email: *wulan_sari@mail.unnes.ac.id.

ABSTRACT

Fashion technology is a compulsory subject taken by students of the Fashion Design education study program. The assessment instruments used in the course have not been developed properly and have not gone through the stages of developing appropriate instruments. This research aims to develop test instruments in the Fashion Technology course that are valid and reliable. This type of research is research and development (R&D). The instruments developed are multiple choice tests and performance tests. Multiple choice test developed as many as 40 questions. The test was tested on 262 students and analysed using ITEMAN 4.0 to determine the difficulty of the items, the discriminating and deceptive power, and the reliability of the instrument. Before being tested, the instrument was first proven its validity by asking for opinions / assessments from 7 experts (expert judgement), and analyzed using the Aiken's V formula<. The results of the analysis of test items concluded 1) the 40 questions developed are valid and feasible to use, have an average of 34.406 and a standard of 9.621 and a minimum score of 4 maximum scores of 37, 2) The level of difficulty (Mean P) of 0.660, is included in the category of questions with a medium level of difficulty, 3) The difference in questions is seen from the Rpbis value of 0, 523, included in the category of very satisfactory differentiation, can distinguish students who have low ability from students who have high ability, 4) Cronbach' Alpha coefficient > 0.7 then it can be interpreted that the instrument has good internal consistency.

Keywords: *Valid and reliable instruments, Test Instruments, Fashion Technology.*

1. INTRODUCTION

Assessment planning and instrument development are very important activities in assessment by educators. Good planning and instrument development can ensure that the assessment is carried out professionally, educationally, and effectively. Assessment planning must be in accordance with the characteristics of competencies, aspects of competencies to be assessed, the level of expected performance achievement, the scope of material, content, and context, how the assessment will be carried out and the time available. While in instrument development, in addition to the above, it is also necessary to pay attention to the rules and principles of assessment. Assessment planning and instrument development are carried out simultaneously during the preparation of semester learning plans (RPS), which are adjusted to the applicable curriculum.

Assessment of learning outcomes by educators aims to monitor and evaluate the process, learning progress, and improvement of student learning outcomes on an ongoing basis. The assessment process involves collecting evidence about student learning achievement, where the evidence is not always obtained through tests alone but can also be collected through observation or self-report[1]. The cognitive process in a person occurs when a person finds a problem, so he can find ideas to solve the problem[2]. Currently, learners are required to think a lot at a higher level according to Bloom's taxonomic theory which emphasizes a higher level of thinking that includes analysis, evaluation, and creation. This level is also called the High order Thingking skill [3], [4].

In educational assessment, the measurement theory used is classical and modern measurement theory. Test scoring in classical measurements is targeted at the time of answering the question items correctly. Scoring is done by giving a score based on the number of correct answers on each item which is then added up to obtain a raw score. Scoring models in classical theory such as this

are considered less relevant to use, because the level of difficulty present at each step tends to be ignored or not taken into account. In analyzing responses to a measurement of higher-order thinking skills, appropriate assessment models are needed, one of which is to use item response theory[5][6].

Assessment in the Fashion Technology course is currently carried out using instruments developed by lecturers who only focus on psychomotor abilities, and cognitive abilities have not been developed properly through proving validity and estimating reliability. So that the instruments used for assessment in the Fashion Technology course are not valid and reliable. In other words, the instrument used has not been able to provide precise and accurate measuring results in accordance with the intended use of the instrument.

Validity is an aspect of measurement accuracy that not only produces precise data, but can also provide a careful picture[7]. Validity can be grouped into three types, namely: (1) criterion-related validity, (2) content validity, and (3) construct validity. The existence of the validity of a test device can be known through analysis of test content and empirical analysis of test scores of item response data[8]. A test is said to be reliable if the observation score has a high correlation with the actual score. Furthermore, it is stated that reliability is the correlation coefficient between two intensity scores obtained from measurement results using parallel tests[9].

Question point analysis is one way to see the quality of the question items or assessments that have been prepared by the question maker, whether the question is feasible or not. The analysis of this question item can be done in a modern or classical way[10]. In the analysis of classic question items, there are three characteristics of question items that can be seen, namely the difficulty of the items, the differentiation of the items, and the function of the deceiver.

The difficulty level of the item is divided into five categories, namely very difficult, difficult, medium, easy, and very easy. The category can be known based on the grain difficulty index. The grain difficulty index ranges from 0 to 1, provided that the category is very difficult (index < 0.1); difficult (index 0.1 - 0.3); medium category (index 0.3–0.7); and easy category (index 0.7–0.9); and very easy category (index > 0.9)[9], [11]. There are various ways to find out the difficulty index of items, but the easiest and most widely used way is to calculate the proportion of answered correctly or proportionally correct (p).

The range of different power indices of question items between values -1 to 1. The meaning of a positive price index describes a testee who is highly able to answer right and a low-ability to answer wrong, vice versa if this index is priced negative. A good question

item is a question item that has a db index of > 0.3. The difference in question items is divided into three categories, namely the good category (db index > 0.3); Reasonably good category (DB index 0.1 – 0.3); and the category is not good (db index < 0.1). The purpose of this study is to develop test instruments in the Fashion Technology course with the help of the ITEMAN 4.3 application.

2. METHOD

This research is development research (R&D), which is used to develop test instruments in the Fashion Technology course in the Fashion Education study program FT UNNES. Instrument development stage[1] The ones used are:

1. Setting assessment goals
2. Arranging the grid
3. Write a problem
4. Study the problem
5. Conduct test trials
6. Analyze question items
7. Fix tests
8. Assembling tests
9. Carrying out the test
10. Interpreting test results

Research instruments before being used to retrieve data, must prove the validity of the instrument and must be used to estimate the consistency and stability of assessment results. Validity refers to the extent to which evidence and theory support the interpretation of test scores required by the proposed use of the test [12]. The validity of the test instrument in the Fashion Technology course is proven through the validity of the content using Aiken's V formula. The aiken formula is used to calculate the content validity coefficient which is based on the results of an assessment from a panel of experts as many as n people on an item in terms of the extent to which the item represents the measured constuk [13]. Aiken's V formula is;

$$V = \Sigma S / [n(c-1)] \tag{1}$$

Information:

V = validity index of Aiken
$S = r - lo$
lo = lowest validity rating number (e.g. 1)
C = highest validity assessment number (e.g. 5)
n = number of raters
r = number given by rater

The results of the item analysis can be said to be valid if they meet the V Aiken limit. Proof of validity on this research instrument using 7 raters and 5 rater scales. Based on table V Aiken[14], so that the limit of the rater coefficient of each grain is 0.75 with a probability of 0.41.

The test instrument was tested on 262 students of the Fashion Education study program FT UNNES. The analysis of the test results is then analysed using ITEMAN. The questions that have been analysed using ITEMAN then look at the characteristics of the difficulty level of the grains, the differentiation power of the grains, and the function of the distractors of each item. We can know these three characteristics by analysing each item based on the results of student responses or answers in answering questions. In addition to seeing these three parameters, ITEMAN can also be used to estimate the reliability of the instrument developed. An instrument is said to be reliable if it has an Alpha Cronbach coefficient of > 0.70[9], [15], [16]. Question items with Cronbach Alpha coefficient < 0.7 are better eliminated and reanalyzed[17].

Table 1. Content Validity Calculation Results.

Item no	S1	S2	S3	S4	S5	S6	S7	Σs	V	Criterion
1	3	4	3	3	3	3	4	23	0.821429	Valid
2	3	4	4	2	3	4	4	24	0.857143	Valid
3	4	3	4	4	3	3	2	23	0.821429	Valid
4	3	4	3	3	3	4	4	24	0.857143	Valid
5	4	3	4	2	4	4	2	23	0.821429	Valid
6	2	2	4	4	4	3	4	23	0.821429	Valid
7	4	3	3	3	4	3	4	24	0.857143	Valid
8	4	2	3	4	4	4	4	25	0.892857	Valid
9	4	4	3	4	4	4	2	25	0.892857	Valid
10	4	4	2	3	3	4	4	24	0.857143	Valid
11	4	2	4	4	3	4	4	25	0.892857	Valid
12	4	3	3	4	4	2	4	24	0.857143	Valid
13	4	4	2	3	4	3	4	24	0.857143	Valid
14	2	3	4	4	4	3	3	23	0.821429	Valid
15	4	2	4	3	3	4	3	23	0.821429	Valid
16	4	4	3	3	3	4	2	23	0.821429	Valid
17	4	3	4	4	3	4	2	24	0.857143	Valid
18	3	2	2	3	1	2	3	16	0.571429	Valid
19	3	4	4	2	4	2	4	23	0.821429	Valid
20	2	4	4	4	2	3	4	23	0.821429	Valid
21	4	3	3	3	4	3	3	23	0.821429	Valid
22	4	3	2	4	4	2	4	23	0.821429	Valid
23	3	4	4	3	3	2	4	23	0.821429	Valid
24	2	4	4	3	4	3	3	23	0.821429	Valid
25	2	3	4	4	4	4	2	23	0.821429	Valid
26	3	2	4	4	4	4	3	24	0.857143	Valid
27	4	2	4	3	4	2	4	23	0.821429	Valid
28	4	2	4	3	3	3	4	23	0.821429	Valid
29	4	4	2	2	4	3	4	23	0.821429	Valid
30	4	3	4	4	3	3	2	23	0.821429	Valid
31	4	3	4	3	4	3	2	23	0.821429	Valid
32	3	2	4	4	4	3	3	23	0.821429	Valid
33	4	4	4	3	2	4	2	23	0.821429	Valid
34	4	4	3	2	3	4	3	23	0.821429	Valid
35	4	3	4	3	3	3	3	23	0.821429	Valid
36	4	3	4	3	4	3	3	24	0.857143	Valid
37	4	2	3	4	4	4	3	24	0.857143	Valid
38	2	4	4	4	4	3	3	24	0.857143	Valid
29	4	3	4	4	3	3	3	24	0.857143	Valid
40	4	4	4	3	3	3	2	23	0.821429	Valid

3. RESULTS AND DISCUSSION

3.1 Content Validity

The test questions for the Fashion Technology course that have been prepared, then proof of the validity of the contents by experts (expert judgement). Content validity is related to rational analysis of the domain to be measured to determine the representation of the instrument with the ability to be measured[18]. Expert judgement examines the material, construction and language / culture used in test items.

From the review conducted by 7 experts, there are several suggestions for improvement in the test items. The test items that have been developed are reassessed by 7 experts which are then proven by content validation. The results of the analysis show that all test items developed are valid, because the material that has been tested is contained in the RPS and is also essential. Proof

of validity on this research instrument using 7 raters and 5 rater scales. Based on Aiken's table V above, the condition for the rater coefficient limit for each item is 0.75 with a probability of 0.41[14]. In Table 1 the following calculation results are presented using the Aiken formula.

3.2 Question Point Analysis

Analysis of the difficulty level of items is divided into three categories, namely easy, medium, and difficult categories. In the 40 items of the Fashion Technology test developed, there were 2 items (5%) included in the difficult item category with a kindergarten index of < 0.3. Question items that have a medium difficulty level of 38 items (95%) with an index of 0.3 – 0.7 and do not have items with easy categories. This can mean that most of the questions are accepted, while the 2 items that fall into the difficult category need to be revised. This is in line with the statement of a good question that is in the category of medium difficulty, which is not too easy or

not too difficult. The good thing is that there are no questions that fall into the category of very difficult or very easy, so there are no question items that are rejected[19].

The difference power index of question items can be used as a consideration to determine whether an item is said to be good or not good. Good means that it can distinguish between high and low ability students, and vice versa, it is not good which means that the grain cannot distinguish high and low ability students. Based on the results in Table 2 shows that out of 38 grains it is said that 95% of the grains have a very satisfactory difference (DB = 0.4 – 1.00). Furthermore, 5% of grains have a satisfactory difference (DB= 0.3 – 0.39). Items that have a db index < 0.1 should be dropped or discarded[11], because the 40 question items developed have a DB > 0.1, there are no question items that need to be revised or discarded.

Table 2. Summary of Question Item Analysis Results.

Item no	Kindergarten	Ket.	DP	Information	Option					Ket.
					A	B	C	D	E	
1	0.671	keep	0.457	very satisfying	0.06	0.55*	0.13	0.06	0.2	good
2	0.536	keep	0.926	very satisfying	0.11	0.079	0.71*	0.05	0.051	good
3	0.476	keep	0.657	very satisfying	0.425*	0.071	0.071	0.121	0.312	good
4	0.387	keep	0.672	very satisfying	0.06	0.12	0.55*	0.13	0.14	good
5	0.623	keep	0.391	Satisfying	0.15	0.23	0.068	0.412*	0.14	good
6	0.523	keep	0.795	very satisfying	0.132	0.131	0.15	0.23	0.357*	good
7	0.671	keep	0.671	very satisfying	0.23	0.068	0.412*	0.14	0.15	good
8	0.636	keep	0.836	very satisfying	0.051	0.11	0.079	0.71*	0.05	good
9	0.476	keep	0.876	very satisfying	0.112	0.325	0.071	0.371*	0.121	good
10	0.387	keep	0.387	very satisfying	0.14	0.06	0.12	0.55*	0.13	good
11	0.623	keep	0.623	very satisfying	0.312	0.425*	0.071	0.071	0.121	good
12	0.523	keep	0.671	very satisfying	0.14	0.06	0.12	0.55*	0.13	good
13	0.356	keep	0.536	very satisfying	0.13	0.06	0.2	0.55*	0.06	good
14	0.622	keep	0.876	very satisfying	0.71*	0.05	0.051	0.079	0.11	good
15	0.567	keep	0.787	very satisfying	0.071	0.121	0.312	0.071	0.425*	good
16	0.243	difficult	0.623	very satisfying	0.55*	0.13	0.14	0.12	0.06	good
17	0.523	keep	0.823	very satisfying	0.068	0.412*	0.14	0.23	0.15	good
18	0.671	keep	0.671	very satisfying	0.079	0.71*	0.051	0.05	0.11	good
19	0.636	keep	0.636	very satisfying	0.071	0.371*	0.112	0.121	0.325	good
20	0.476	keep	0.776	very satisfying	0.12	0.55*	0.14	0.13	0.06	good
21	0.387	keep	0.387	Satisfying	0.071	0.071	0.312	0.121	0.425*	good
22	0.623	keep	0.623	very satisfying	0.12	0.55*	0.14	0.13	0.06	good
23	0.523	keep	0.523	very satisfying	0.2	0.55*	0.13	0.06	0.06	good
24	0.356	keep	0.556	very satisfying	0.079	0.71*	0.05	0.051	0.11	good
25	0.457	keep	0.622	very satisfying	0.071	0.071	0.121	0.312	0.425	good
26	0.526	keep	0.567	very satisfying	0.12	0.55*	0.13	0.14	0.06	good
27	0.657	keep	0.943	very satisfying	0.23	0.068	0.412*	0.14	0.15	good
28	0.672	keep	0.823	very satisfying	0.131	0.15	0.23	0.357*	0.132	good
29	0.391	keep	0.667	very satisfying	0.068	0.412*	0.14	0.15	0.23	good

Item no	Kindergarten	Ket.	DP	Information	Option					Ket.
					A	B	C	D	E	
30	0.295	difficult	0.943	very satisfying	0.11	0.079	0.71*	0.05	0.051	good
31	0.671	keep	0.523	very satisfying	0.325	0.071	0.371*	0.121	0.112	good
32	0.636	keep	0.671	very satisfying	0.06	0.12	0.55*	0.13	0.14	good
33	0.476	keep	0.636	very satisfying	0.425*	0.071	0.071	0.121	0.312	good
34	0.387	keep	0.776	very satisfying	0.06	0.12	0.55*	0.13	0.14	good
35	0.623	keep	0.887	very satisfying	0.06	0.2	0.55*	0.06	0.13	good
36	0.573	keep	0.623	very satisfying	0.05	0.051	0.079	0.11	0.71*	good
37	0.356	keep	0.723	very satisfying	0.121	0.312	0.071	0.425*	0.071	good
38	0.457	keep	0.856	very satisfying	0.131	0.132	0.23	0.357*	0.15	good
39	0.421	keep	0.823	very satisfying	0.068	0.23	0.14	0.15	0.412*	good
40	0.326	keep	0.767	very satisfying	0.11	0.051	0.71*	0.05	0.079	good

Table 3. Statistical Summary.

Score	Items	Mean	SD	Min Score	Max Score	Mean P	Mean Rpbis
All items	40	34.406	9.621	4	37	0.66	0.523
Scored Items	40	34.406	9.621	4	37	0.66	0.523
1	4	3.719	7.553	2	4	0.557	0.732
2	8	6.688	4.206	4	8	0.669	0.697
3	4	2.1	2.621	2	4	0.458	0.785
4	6	4.49	3.621	3	6	0.674	0.689
5	8	5.97	7.553	5	8	0.572	0.876
6	5	3.236	2.206	3	5	0.678	0.768
7	3	2.005	1.553	2	3	0.62	0.895
8	2	1.27	1.206	0	2	0.617	0.739
Scaled Total	40	0	1	-3.16	0.981	-	-
Scaled 1	4	0	1	-3.14	0.567	-	-
Scaled 2	8	0	1	-3.031	0.795	-	-
Scaled 3	4	0	1	-3.19	0.571	-	-
Scaled 4	6	0	1	-3.14	0.967	-	-
Scaled 5	8	0	1	-3.031	0.595	-	-
Scaled 6	5	0	1	-3.14	0.781	-	-
Scaled 7	3	0	1	-3.64	0.567	-	-
Scaled 8	2	0	1	-3.231	0.895	-	-

Based on the table above, it can be seen that there are 40 questions analyzed and an average of 34,406 and a standard of 9, 621 and a minimum score of 4 with a maximum score of 37. The difficulty level (Mean P) of 0.660, belongs to the category of questions with a medium level of difficulty. The difference in questions seen from the Rpbis value of 0.523, included in the category of very satisfactory difference power, can distinguish students who have low abilities from students who have high abilities.

3.3 Reliability

Based on the reliability table, the value of Cronbach's alpha reliability coefficient for the entire instrument is 0.975, the basic stitch material is 0.970, the potency material is 0.882, the finishing material is 0.870, the pleated material is 0.872, the belahan material is 0.870, the pocket material is 0.892, the sleeve material is 0.857 and the buttonhole material is 0.882. The standard measurement error (SEM) for the entire instrument is 1.512. While SEM on sewing basic puncture material is 1.302, potent material is 0.757, clothing finishing material is 1.402, pleated material is 0.657, belahan material is 1.372, pocket material is 0.857, sleeves material is 1.502 and buttonhole material is 0.787. Since Cronbach' Alpha coefficient > 0.7, it can be interpreted that the instrument has good internal consistency.

Table 4. Reliability Calculation Results.

Score	Alpha	SEM	Split-Half (Random)	Split-Half (First-Last)	Split-Half (Odd-Even)	S-B Random	S-B First-Last	S-B Odd-Even
Scored items	0.975	1.512	0.944	0.966	0.937	0.971	0.983	0.968
1 (basic stitching skewer)	0.97	1.302	0.939	0.97	0.925	0.969	0.985	0.961
2 (kampuh)	0.882	0.757	0.721	0.793	0.759	0.838	0.884	0.863
3 (penyelesaian tepi busana)	0.87	1.402	0.946	0.87	0.925	0.969	0.985	0.838
4 (pleated)	0.872	0.657	0.761	0.797	0.729	0.985	0.838	0.985
5 (belahan)	0.87	1.372	0.839	0.88	0.825	0.884	0.969	0.961
6 (pocket)	0.892	0.857	0.721	0.763	0.754	0.838	0.838	0.969
7 (sleeves)	0.857	1.502	0.972	0.87	0.725	0.969	0.985	0.961
8 (buttonhole)	0.882	0.787	0.821	0.723	0.755	0.838	0.884	0.863

4. CONCLUSION

When viewed from the results of statistical item analysis, it can be seen that question item no. 1 has a key error, so it needs to be revised. Question items no. 2 – 40 have good question criteria and are worth using. However, when viewed from the statistical option, it can be seen that all questions need to be reviewed / revised because there are several alternative answers that do not function as a distractor / deceiver and cannot be used to distinguish the ability of test takers. There are many question items with Rpbis scores on negative answer alternatives, which means that test takers who have high abilities tend to choose alternative answers to deceivers (not keys). And conversely, test takers who have low ability tend not to choose alternative deceptive answers (not keys).

REFERENCES

[1] D. Mardapi, Measurement, Assessment and Evaluation of Education, 2nd ed. Yogyakarta: Parama Publishing, 2019.

[2] A. Klieger and G. Sherman, Physics Textbooks: Do They Promote or Inhibit Students' Creative Thinking, Phys. Educ., 50(3), 2015, pp. 305–309. DOI 10.1088/0031-9120/50/3/305

[3] I. Basuki and Hariyanto, Learning Assessment. Bandung: PT Remaja Rosdakarya, 2014.

[4] E. Istiyono, D. Mardapi, and Suparno, "Development of high-level thinking skills tests for physics (pysthots) of high school students," J. Researcher. and Eval. Educators., Vol. 18, No. 1, 2014, pp. 1–12. Available: https://doi.org/10.21831/pep.v18i1.2120.

[5] F. M. Arshid, J. N. Desai, D. J. Duff, C. H. Giles, S. K. Jain, and I. R. Macnral, "A Study of Certain Natural Dyes I-The Adsorption of Brazilwood and Logwood Colouring Matters by Fibres," J. Soc. Dye. Colour., Vol. 70, No. 9, Sep. 1954, pp. 392–401, DOI: 10.1111/J.1478-4408.1954.TB02043.X.

[6] V. der Linden, W. J, and R. K. Hambleton, Handbook of Modren Item Response Theory. New York: Springer-Verlag New York, Inc., 1997.

[7] E. Istiyono, Development of Assessment Instruments and Analysis of Physics Learning Outcomes with Classical and Modern Test Theory. Yogyakarta: UNY Press, 2020.

[8] W. Lissitz and K. Samuelsen, "Further clarification regarding validity and education," Educ. Res., Vol. 36, No. 8, 2007, pp. 482–484.

[9] M. J. Allen and W. M. Yen, Introduction to Measurement Theory. Texas: McGraw-Hill, 1979.

[10] D. M.-Y. UNY and undefined 2010, "Character Education Assessment," academia.edu, Available: https://www.academia.edu/download/32033941/Penilaian_karakter.pdf.

[11] S. Azwar, Achievement Test: Learning Achievement Measurement Development Function. Yogyakarta: Learning Library, 2016.

[12] AERA, APA, and NCME, The Standards for Educational and Psychological Testing. 2014.

[13] L. R. Aiken, Psychological Testing and Assessment, 11th ed. 2003.

[14] L. R. Aiken, "Three Coefficients for Analyzing the Reliability and Validity of Ratings:," http://dx.doi.org/10.1177/0013164485451012, Vol. 45, No. 1, Sep. 2016, pp. 131–142, DOI: 10.1177/0013164485451012.

[15] R. L. Linn, Educational Measurement. New York: Mc Millan Publishing, 1989.

[16] D. L. Streiner, "Starting at The Beginning: An Introduction to Coefficient Alpha and Internal

Consistency," J. Press. Assess., Vol. 80, No. 1, 2003, pp. 99–103.

[17] M. F. Bahri and Supahar, "Content Validity and Reliability Analysis of Integrated Islamic Science Test Instrument to Measure The Student's Critical Thinking Ability," J. Islam. Soc. Stud., Vol. 5, No. 1, 2019.

[18] H. Retnawati, Quantitative Analysis of Research Instruments (Researcher, Student, and Psychometric Guide). Yogyakarta: Parama Publishing, 2016.

[19] A. N. Amalia and A. Widayati, "Analysis of class XII high school quality control test questions for accounting economics subjects in Yogyakarta City in 2012," J. Educator. Accountant. Indones., Vol. 10, No. 1, 2012.