CHAPTER 5

VALIDITY STATISTICAL ANALYSIS

Validity is defined as the extent to which measurements are useful in making decisions and providing explanations relevant to a given purpose (Sax 1997:304). The 1985 standards for educational and psychological testing define validity as "the appropriateness, meaningfulness, and usefulness of the specific inference made from the test score". The importance of a measuring instrument to be valid is that the score of the measurement is used to make a judgment on the progress of grading of a student. Validity can be divided into external validity and internal validity. According to Cook and Campbell (1979), external validity refers to how far is it possible to generalise the results from the sample of the population. Internal validity is traditionally related to the question: Are we measuring the thing we were supposed to measure? For the purpose of this study when referring to validity, internal validity is considered. So how would one go about determining the validity of the inference done from a test?

A test in itself is valid if it does in fact measure what it claims to measure. Tredoux and Durrheim (2002:216) state that this is not an easy judgment to make, as there is no direct measure of validity. In general, this judgment depends on whether the test leads to inferences that are meaningful and useful. Lloyd-Jones (1986) has a more educational approach to validity and frame validity around two questions: *Firstly*, is what you, the educator, expect of your students to learn justifiable and reasonable? *Secondly*, do the methods of assessment achieve what they set out to do? In their view, affirmative answers to the questions would validate the form of assessment.

Oosterhof (1994:53) states that validity pertains to the degree to which a test measures what it is supposed to measure. This view is not much different from those of all the other experts in educational measurement. Oosterhof (1994:53) states: "More than any other factor, the quality of a test depends on its validity." He elaborates by saying that if a test does not measure what it is supposed to measure, it is useless. These strong words place a heavy burden on the shoulders of educators to make sure that tests compiled are valid at all times. This is easier said than done, therefore there is a need for the CCAT tool as

described in Chapter 2 to assist educators in striving to administer tests to students that will be valid, reliable and fair (see 2.5).

Apart from the CCAT tool, the educator or test constructor must be assisted by the assessment committee and follow the 'Valid, Reliable and Fair assessment model' as proposed in the conclusion (Chapter 6) to ensure validity, reliability and fairness of test score interpretations. If a test is not valid, it then implies that it can definitely not be fair because it is not measuring what it is supposed to measure, and the students did not prepare for it, since they were expecting something else. Validity is a complex concept and therefore an extensive discussion of validity is necessary. What follows is a discussion of the types of validity as agreed upon by many experts in the field of educational measurements such as Ebel (1979), Messick and Ross (1962) and Thorndike (1961). A schematic presentation of the different components of validity are outlined in Figure 24 and subsequently discussed.

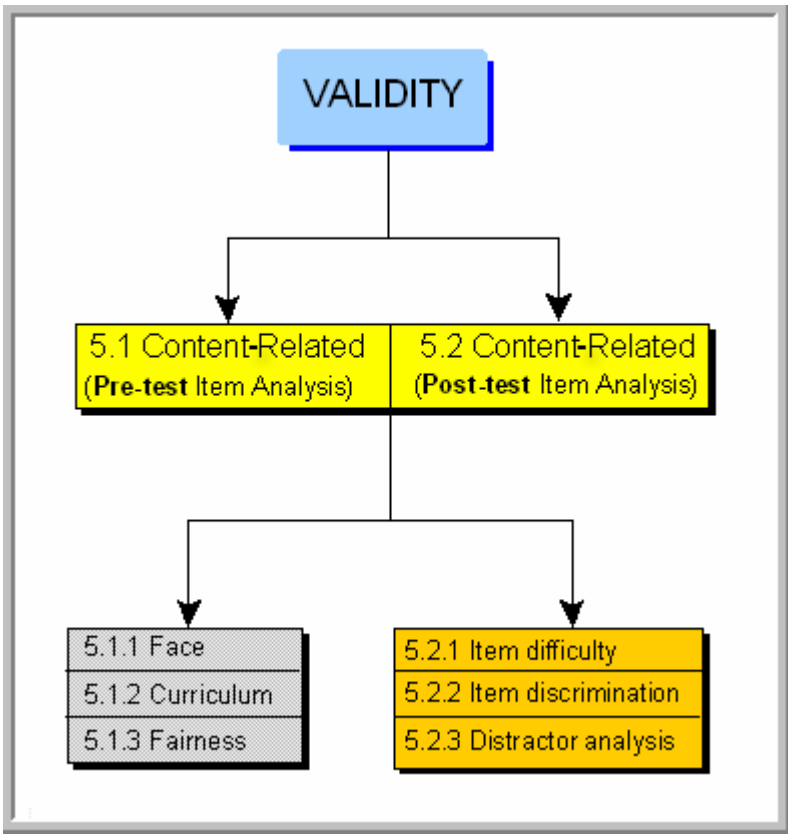

Figure 24 Validity

5.1. CONTENT-RELATED (Pre-test item analysis)

Content-related validity refers to how well the test items represent the domain of the construct being measured as defined by statisticians (Tredoux & Durrheim 2002:217). Sax (1997:305) defines content validity as the extent to which items both measure specific objectives and reflect the universe, or domain of tasks, under consideration. In the aforementioned definition by Sax (1997) the reference to the fact that items must measure specific objectives is worth noting. This is an important reason why all the banked items in a CCAT must be referenced to the educational outcomes (see 6.4) specified for the construct being measured, and why the CCAT must have the facility to retrieve items by their outcomes. The items must adequately represent the construct proposed to be measured. Assessment focuses on what was taught and emphasised and the assessment needs to be clearly related to the content of the course presented. There are two ways to ascertain the content validity of the test **before** the test is administered, namely face validity and curriculum validity. Item analysis is used to ascertain content validity **after** the test has been administered and all these results are used to improve the quality of the item which in turn will improve the overall content validity of the test. Each of these procedures will now be discussed under subsequent headings.

5.1.1 Face validity

Discussion

Face validity refers to the appearance of the test (Tredoux & Durrheim 2002:217). It is very important that the test always appear authentic to the students taking the test. If students are, for example, told that the purpose of the test is to only test certain outcomes from a domain, but the test includes items from other domains, they might feel that they have been deceived. While face validity is not strictly a criterion for validity, it does have an effect on the test scores if the participants have doubts about the test (Tredoux & Durrheim 2002:217). Face validity is considered to be a desirable quality within test, but measurements experts do not consider this quality basic to the validity of a test (Oosterhof 1994:60).

Results obtained

The attached test in Appendix A was submitted to the assessment committee to ensure face validity.

5.1.2 Curriculum validity

Discussion

Another term that has become associated with validity is *curricular validity.* Hills (1981:61) define curricular validity as "an evaluation of the extent to which the content of a test agrees with the content of instruction". Often curricular validity is subdivided into issues of curricular and instructional validity. When used as distinct concepts, curricular validity pertains to the content covered in curricular material (syllabus), whereas instructional validity refers to the degree that this content is actually taught to students (Oosterhof 1994:60). Put together, this implies that curriculum validity concerns the agreement amongst students and educators about test content with the instructional content, so that it ties up with the set educational outcomes. What method can be put in place to ensure that instructional content ties up with the set educational outcomes?

A method to ensure that students and educators agree on test content is to make sure that a comprehensive learning guide is in place. The learning guide must therefore clearly indicate the curriculum outcomes holistically, as well as critical outcomes and the assessment criteria. Both the students and educators can use the learning guide as the agreement on the curriculum and the instruction that needs to take place to enhance learning and ensure the validity of the test.

There needs to be a clear idea of the expected learning outcomes and these outcomes must be justifiable in terms of the curriculum. Evidence of curricular validity is obtained by determining the degree of incongruence or mismatch. This is based on a systematic, judgmental review of the test against the curricular objectives or materials by content

experts, also referred to as Subject Matter Experts (SMEs). These experts may be classroom educators or curriculum specialists; they are the only professionals in a position to judge curricular validity (Berk 1986:116). They independently examine the items and decide whether each of the items is weakly relevant or strongly relevant to the content domain of the construct (Tredoux & Durrheim 2002:217). In most cases, curricular validity must preferably take place before a test is administered to students.

Content-related validity is a direct result of the items used in a measurement instrument. The quality of the items used will have a direct influence on the content validity, therefore they need to be carefully reviewed and used carefully. One cannot hope to perfect items to the point where a hypercritical reviewer cannot quibble over conceivable ambiguities or exceptions of the keyed answers (Cronbach 1973:457), but there is nothing wrong with trying to perfect items with the implementation of procedures to do so. Obviously, the more complex the subject matter, the more it is open to some misinterpretation. The Assessment Committee (2006) were asked to do a critical evaluation of the test items in terms of their relevance to curriculum (content) validity. This again provides a strong motivation for the implementation of an assessment committee at a departmental level (see 2.7.3.1). Following are the results of the feedback obtained from the relevant Assessment Committee (2006).

Results obtained from pre-test item analysis

Tredoux and Durrheim (2002:217) provide a simple formula for empirically 'measuring' the extent of content validity by calculating the proportion (or percentage) of items that the subject experts agreed were strongly relevant. This measure ranges from 0 to 1.00 (0% to 100%).

$$\text{Content validity} = \frac{x}{N}$$

where:
x = number of items evaluated as strongly relevant by both judges
N = total number of items in the test

The Assessment Committee (2006) rated 51 items average as strongly relevant out of the 54 items asked. This gives a result of 0.95, which indicates high content validity. A pre-test item analysis by the Assessment Committee (2006) as suggested in 3.2 revealed that the following items could be ambiguous. The Assessment Committee (2006) made the following suggestions for improvement:

- With respect to Item 13 as indicated in Figure 25, it was suggested that the type of flip-flop be specified as part of the wording in the item as there are various types of flip-flops available, and the students might be uninformed if this detail is omitted.

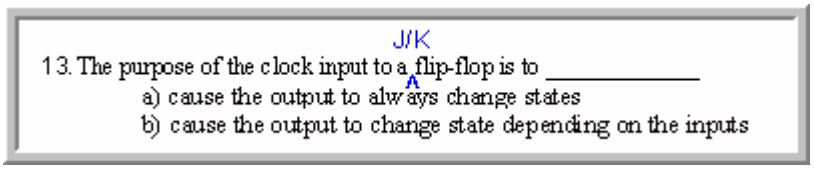

Figure 25 Vague wording in an item

- In Figure 26, the wording of this item was considered to be "slightly confusing" and it was suggested that the polarity of the D flip-flop be specified as part of the wording in the item. Another problem with this item was that the wording of the two distractors, b) and c), is such that both could be considered as correct, but the answer sheet indicated b) to be correct. The wording of c) needed to be changed. No suggestions were given. The typing mistake was changed from *fl* to *flip*.

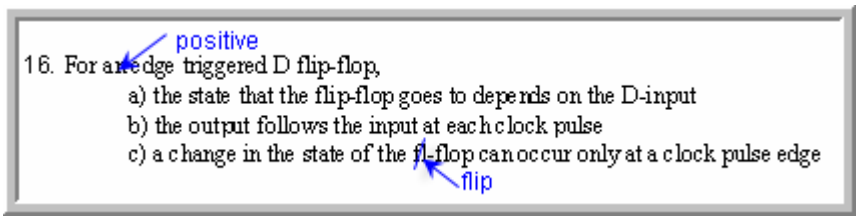

Figure 26 Item with incorrect wording

- It was suggested that "more possible distractors" to the item be added as this will lead to an improvement of this item, as the addition of distractors enhances the quality of the item, because they are also possible answers to the item. The addition of distractors is shown in Figure 27.

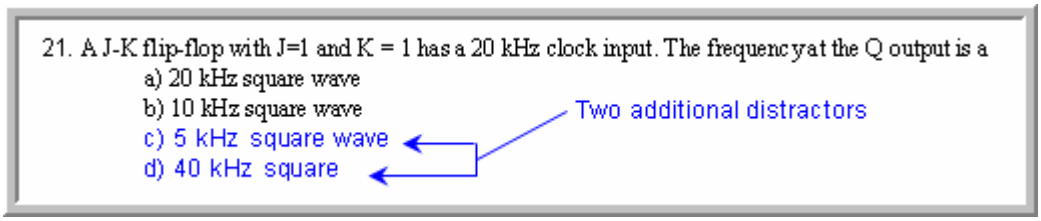

Figure 27 Item with inadequate distractors

- The item in Figure 28 tries to establish the students' knowledge of two different types of flip-flops, but the other type of flip-flop is not pertinently mentioned, so the item could be ambiguous to the student and the inference made from the result of this item could then not be very accurate.

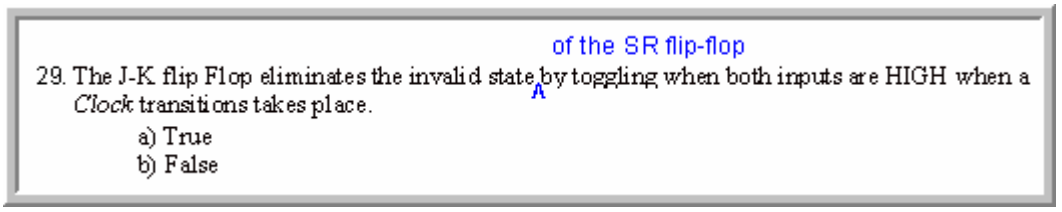

Figure 28 Ambiguous item

The Assessment Committee (2006) further found that the test focuses primarily on knowledge and to a lesser extent on comprehension. It was therefore judged to be better suited for formative than for summative assessment. Improvements were suggested for Items 4 and 9. Items 13 and 29 were identified as possibly ambiguous. Typing/spelling mistakes were noted in Items 16 and 35. There was one item (14) for which an incorrect answer was given on the answer sheet. In a few instances Items 2, 3, 14, and 31 appeared to be repetitions of Items 5, 7, 18 and 37 respectively. The wording was different, but the knowledge tested was the same. This may have been deliberate or not but it was well worth noting.

5.1.3 Fairness

Discussion

Assessments which claim to be highly valid and reliable but discriminate against members of any ethnic/racial or gender group or against students with disabilities should not be used

because they are not fair to those students and are therefore not valid (Metsämuuronen 2002:20).

Educators must at all times ensure that assessment procedures are fair to all students, and that they do not discriminate against members of any ethnic, racial or gender group or against students with disabilities (Metsämuuronen 2002:20). Salvia and Ysseldyke (2001) describes fairness as an imprecise concept both psychometrically and legally and is best viewed as a marker for a class of conditions and situations in which the outcomes are thought to be disadvantageous, inaccurate, or wrong. An important fact that Salvia and Ysseldyke (2001) mention is that unfairness usually implies dissatisfaction with an outcome. HE institutions are becoming more and more accountable for the outputs that they deliver in terms of students passing. Outputs are directly linked to the testing and grading of students. Students can lodge allegations of unfair procedures which might focus on any of the following complaints as listed by Salvia and Ysseldyke (2001:39):

- Students may lack opportunity, possibly as a result of inadequate resources, inadequate instruction, student deficiencies or inadequate home supervision.
- It could be possible for students to disagree with the grading they received. In particular, if it reflects a fail score, this could even lead to court cases, which must be avoided at all times. This could have disastrous impact on the assessment of the institution and the qualifications it issues. Procedures must be in place to ensure that items used in tests are examined for fairness.

The best guarantee of fairness is a much more comprehensive approach to quality assurance. Educators and students will work in a context where checks on quality, standards and professional consistency are routine (Pahad 1997:51).

Results obtained

Any inconsistency in an item will affect the fairness of the item. There is no single calculation that can be used to determine the fairness of an item. However, the role of the assessment committee should be to ensure fairness. This can be complemented by

statistical measurements.

Structured interviews were conducted with three staff members who used the CCAT. Some of the significant responses are presented verbatim below:

- *Record management is much easier.*
- *The establishment of a standardised moderated question bank seemed ambitious at first, but with the reliability calculator, I can really distinguish between good and bad questions.*
- *More assessment opportunities on smaller chunks of work are positively perceived by the students.*
- *Although multiple choice, the same concept could be asked from different angles, turning assessment into a learning experience for students.*
- *Students' attention was captivated by the use of computers for assessment*.
- *Invigilation is simplified by the fact that questions and / or answers can be shuffled, reducing the risk of plagiarism (copying from other students).*
- *Item analysis of items also turned the assessment into a learning experience for me as an educator.*
- *The flexibility in calculating a final score for a student is really beneficial, making continuous assessment a feasible assessment approach.*
- *Less marking leads to more time for subject development.*
- *Although marking is still required for essay questions, certain components of the work were effectively covered by using multiple choices.*
- *The possibility of inserting images as distractors enabled me to ask more understanding and analysis items as students do not have to draw the circuits.*

5.2 CONTENT-RELATED (Post-test item analysis)

Discussion

As mentioned in 1.2, content-related validity is a direct result of the items used in a measurement instrument and no test item is perfect, no matter how much time and

consideration it receives (Sax 1997:236). Even the most carefully prepared items and those that have been pre-test checked by the Assessment Committee (2006) are susceptible to human error and on analysis may prove to be ambiguous to students, too simple, overly difficult, or non-discriminating (Sax 1997:236). Another important factor in doing item analysis is helping the educator detect student misunderstanding of items and how they respond to items. All of this feeds back to improve the overall validity of the inference made from the measurement instrument we use. So a pre-test item analysis by the Assessment Committee (2006) is considered not to be sufficient in establishing overall validity of the inference made from the scores of a test.

Running an item analysis after a test has been administered to students and scored is as much part of the educator's responsibility as is selecting appropriate teaching aids, and developing an effective lesson for discussion (Sax 1997:236). All of these activities are designed to improve instruction by helping the educator to make the most effective decisions for the students. In doing this educators can use this information gathered to develop more effective measuring instruments. Item analysis is particularly effective for objective items, but it can also be applied just as effectively for other item formats (Sax 1997:236). A test composed of items revised and selected on the basis of item analysis is almost certain to be much more valid than one composed of an equal number of untested items (Ebel 1979:258). When item analysis is done properly it can serve the following purposes as summarised from Sax (1997:236):

- It enhances the technical quality of a test or examination.
- It facilitates instruction, that is, determine student weakness in a certain area of learning.
- It assists educators in setting better examinations.
- It helps to save educators' time in the long run as they can reuse items that have been analysed and banked.

A new emphasis in the development of tests is on the establishment of the validity of computerised interpretations and computer-generated reports. The developments of specifications for a fully interpretive computer report force educators to pose detailed

questions relating the research base of a test to specific numerical rules of score interpretation (Roid 1989:35). Despite the importance of computer analysis during the development of tests, there remains a surprising void in the availability of computer programs specifically designed in an integrated package for the development and analysis of tests and test items. To be able to develop a comprehensive computerised testing instrument would require a person simultaneously skilled in the content area, computer programming, and measurement statistics. If one can put together a team of experts encompassing all these, is it possible to develop a tool that will comply with all the requirements of a Comprehensive Computerised Assessment Tool. Examples are subsequently provided of statistical reports that are generated by the CCAT tool to assist educators in making informed decisions on the tests that have been constructed and the interpretation of the scores obtained after the tests have been administered to the students.

5.2.1 Item difficulty

Discussion

The difficulty of an item refers to the proportion of students who answered the item correctly, expressed as a percentage. To determine the difficulty level of individual test items, the number of students who answered an item correctly is divided by the total number of students who answered the item. Multiply that figure by 100. Values for the difficulty index range from 0% (very difficult) to 100% (very easy).

What are these values used for?

The value obtained from the difficulty index is used to decide which items need to be analysed. Items answered correctly by the majority of the students need to be examined for clues within the item, whether they be grammatical or whether the answers could have been derived from other items in the same test. On the other hand, if most of the students answer an item incorrectly, it could well be that the item is ambiguous or confusing. It could also indicate that the domain tested was not adequately covered during instruction. If the purpose for the test is to grade students, items with a difficulty level around 50% will

be used. However, if Bloom's taxonomy is used, then it is impossible not to have items that vary in difficulty.

Results obtained

The following figure shows part of a difficulty report provided by the CCAT, listing the difficulty factor for all items used in a test. Item 8 needs to be analysed because of its high difficulty level.

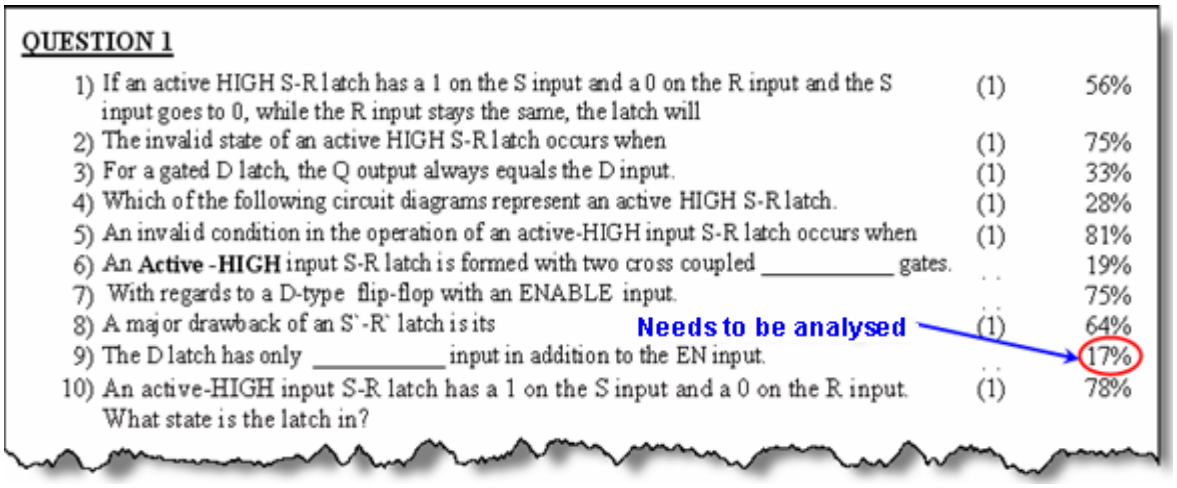

Figure 29 Sample report of difficulty factor for all items

5.2.2 Item discrimination

Discussion

One of the most powerful indicators of an item's quality is the item discriminating index. In brief, an item discrimination index typically indicates how frequently an item is answered correctly by those who perform well on the total test (Popham 1995:201). One approach to computing an item discrimination statistic is to calculate a correlation coefficient between students' total test scores and their performance on a particular item. A correlation is a statistic that quantifies the relationship between two variables. To be able to correlate a test item, a categorical variable (i.e., the student either answered the test item correctly or incorrectly), with a continuous variable (i.e., percent score on the

examination), the correlation index for the categorical variable and the continuous variable is needed. This is the point biseral correlation. The point biseral ranges from -1.00 to + 1.00. What is a desirable point biseral correlation for a test item? The higher the better. As a general rule, +20 is desirable (Tredoux & Durrheim 2002:220). However, there is an interaction between item discrimination and item difficulty, and one should be aware of two major factors:

- Very easy or very difficult test items have little discrimination.
- Items of moderate difficulty (60% to 80% answering correctly) generally are more discriminating.

$$IDsI = \frac{T - B}{N}$$

where:
T = number of the top 25% that correctly answered the item
B = number of the bottom 25% that correctly answered the item
N = the total number of people in either the top or bottom 25%

It is important to recall that the primary role of a test is to help the educator distinguish between students who need further instruction and students who have become proficient in a particular set of skills (Oosterhof 1994:199). For that reason, test questions that have a high item discrimination index are desired.

A test item, by itself, is usually not a very reliable measure of a student's achievement. This is one reason why a test must consist of numerous items. This is also the reason why one should not expect an individual item to have a discrimination index that is too high. Items that are of moderate difficulty; that are neither extremely easy nor extremely difficult, are likely to obtain a high level of discrimination. For maximum discrimination, item difficulty should be somewhere in the range of 0.5 to 0.9, preferably within the middle to lower part of the range (Oosterhof 1994:199).

A positively discriminating item indicates that an item is answered correctly more often by those who score well on the total test than by those who score poorly on the total test. A negatively discriminating item is answered correctly more often by those who score poorly on the total test than by those who score well on the total test. A non-discriminating item is one for which there is no substantial difference in the correct response proportions of those

who score well or poorly on the total test. The verdict from all of this is that educators would prefer that their items are positive discriminators, because positively discriminating items tend to be answered correctly by almost all knowledgeable students (those who scored high on the total test) and incorrectly by the least knowledgeable students (those who scored low on the total test). Negatively discriminating items indicate that something is wrong, because the items tend to be missed more often by the most knowledgeable students and answered correctly more often by the least knowledgeable students. If the primary goal of item selection is to maximise test reliability, as it probably should be for most classroom tests, the items with the highest discrimination in terms of the index should be chosen (Ebel 1979:63).

Item discrimination indices of all types are subject to considerable sampling error (Pyrczak 1973). The smaller the sample of students' answer sheets on an item used in the analysis, the larger the sampling errors. The results obtained from achievement tests are highly dependent on the type of instruction that the students received relative to the item of which the discriminating index is being determined. But even though one cannot determine the discriminating indices of individual items reliably without using large samples of student responses, item analysis based on small samples is still worthwhile as a means of overall test improvement (Ebel 1979:63). The question now arises: What should an items discriminating index be for the educator to consider an item to be acceptable for inclusion in a test? Ebel (1979) offers the experienced-based guidelines as shown in Table 2.

Table 2 Guidelines for evaluating the discriminating efficiency of items

Discrimination index	**Item evaluation**
0.4 and above	Very good items
0.3 – 0.39	Reasonably good items. but possibly subject to improvement
0.2 – 0.29	Marginal items, usually needing improvement
0.19 and below	Poor items, to be rejected or improved by revision

Designers of norm-referenced tests typically seek items in the range of .35 to .70. If most students in both the high and the low groups respond to an item correctly, the discrimination index might be .14. This is a red flag that the test item is too easy. If more

students in the low group respond to an item correctly than students in the high group, the discrimination index will be negative (-.08). This is a red flag that the test item is flawed.

In criterion-referenced tests there need not to be as much discrimination as in the case of norm-referenced tests, as criterion-referenced tests gauge the mastery level of students. In particular, when a specific part of instruction is being measured, it is not abnormal for most of the students to obtain a high score in most of the items. Criterion referenced tests are not used for summative or grading purposes, that is where norm-referenced tests are used.

Results obtained

An item analysis report can be used to identify items for which the answer sheet has a wrong answer as the correct answer on the answer sheet. Figure 30 shows item 14 with a difficulty factor of 0.3077 indicating a difficult item. On closer analysis it was established that the answer sheet indicated a wrong answer as correct. Analysing the graph in Figure 32 item 14 is clearly a fall out, compared to the other items in the test.

14) A positive edge triggered flip-flop changes state with a transition on the clock input.	Upper Quartile	Lower Quartile	Count	%	Disc Index
a) HIGH-to-HIGH	2	3	7	16%	-0.091
b) LOW-to-LOW	0	0	0	0%	0.000
c) LOW-to-HIGH ← Correct answer →	8	4	21	50%	0.364
d) HIGH-to-LOW ← Incorrect answer →	2	4	13	30%	-0.182
Marks: (1) Percentage Correct: 28% Difficulty Factor: 0.3077					

Figure 30 Item with wrong answer as the correct answer on the answer sheet

In Figure 31, the questions and possible answers for Item 50 (as it was numbered in the test) are displayed. The question number 635 as shown in the figure is the number of the item in the question bank.

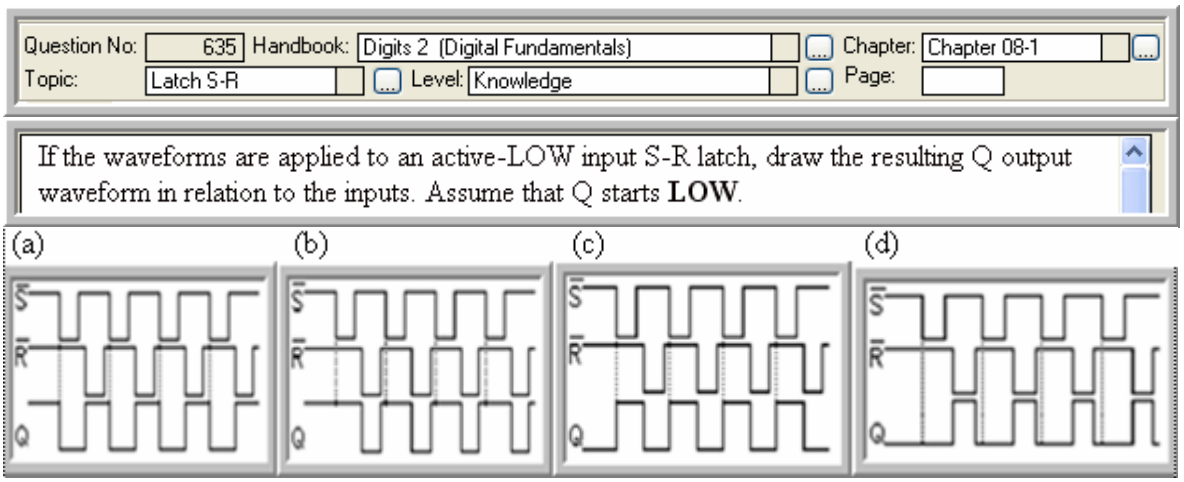

Figure 31 Item with an obvious clue for the students

The problem with this question is that it states that the Q output starts LOW. On analysing the options for the item it can be seen that in distractors (a) and (b) the Q waveform starts HIGH. This is an obvious clue for the students. This is more likely a mistake and not deliberate distractors. Nevertheless, on analysing the students' results most definitely grasped that it was a clue. Two deductions can be made from this. Firstly, a mistake such as this would not have been there if the item had been properly moderated, and secondly if they had been meant to be first-class distractors, then the results from the item analysis clearly indicate that they were not good distractors.

The solution to this question is apparent. Firstly, the question could be fixed by removing the part "Assume that Q starts **LOW**". The subject specialist will know that this is not the best fix, because all of these types of questions usually indicate the initial value of the Q output to the students. Therefore the second fix would be the most appropriate namely that the two distractors must be fixed. An item analysis of the student responses listed in Table 1, revealed the following (see Table 3):

Table 3 Item analysis of student responses

Item 50	Upper Quartile	Lower Quartile	Count	Percentage	Disc Index
(a)	0	1	3	8%	-0.100
(b)	0	4	5	13%	-0.400
(c)	4	0	8	22%	0.400
(d)	6	4	20	55	0.200

From Table 3 it can be seen that none of the better students (those in the upper quartile), selected the faulty distractors (a) or (b). Figure 32 shows the advantage of using a graph that highlights the faulty items.

Item correlation

For each item the primary indicator of its power to discriminate amongst students is the correlation coefficient. The correlation coefficient reflects the tendency that students who select the correct answer also have high scores in the test.

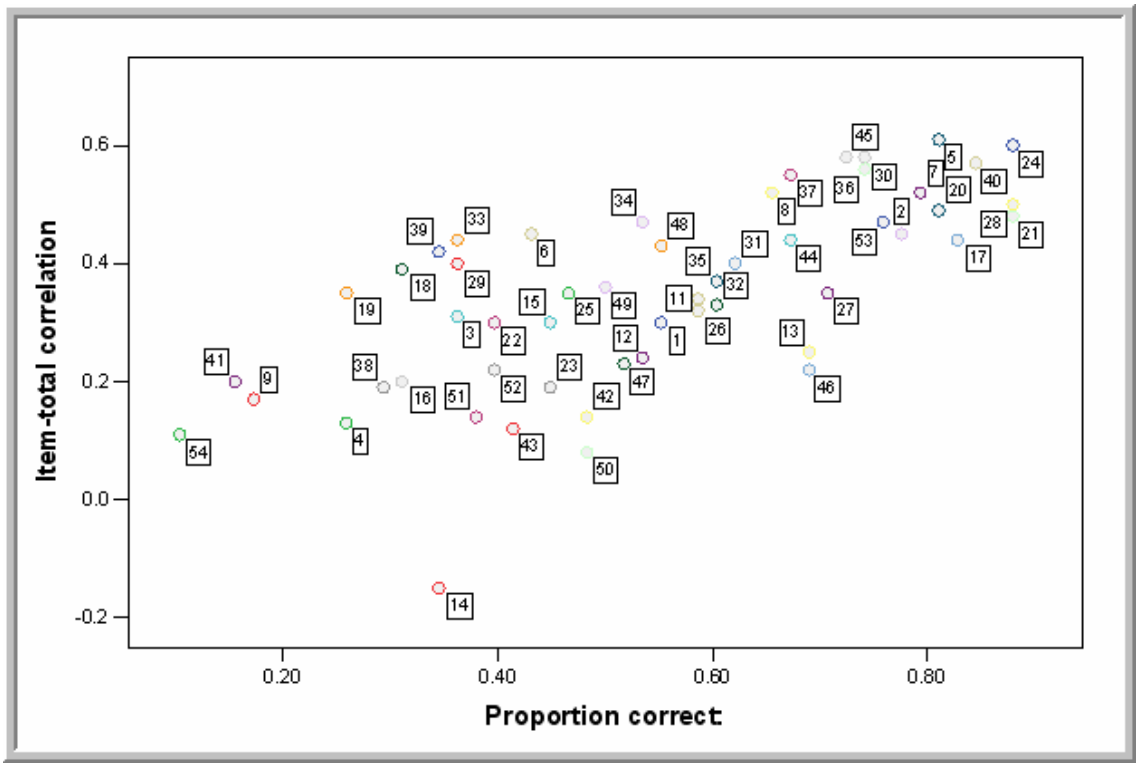

Figure 32 Classification of items and item-total correlation

In items that have been answered correctly or incorrectly by a large proportion, more than 85% of the students have a noticeably limited ability to discriminate. To consider a test as good, most items should be answered correctly by 30% to 80% of the students.

5.2.3 Distractor analysis

Discussion

The importance of distractor analysis lies in the fact that all distractors should contribute to the difficulty or discrimination of an item. If a distractor has a low selection or no selection frequency then it should be revised or discarded entirely and be replaced by another distractor. The effectiveness of a distractor can only be determined after being administered to students again. When doing a pre-test item analysis as described in 5.1.2, just as much attention must be given to distractors as to the item itself.

What should we expect from the results obtained for an item?

For a test item to perform satisfactorily, a greater number of students in the upper quartile should answer the item correctly than those in the lower quartile. It is expected that a smaller number of students in the upper quartile should select the wrong distractor to the item (Oosterhof 1994:200). Item distractors must be analysed to confirm that they do perform as expected. Under the heading "Results obtained", examples of some items are indicated in Figures 33 (a) and (b), where distractors are discussed.

Results obtained

2) The invalid state of an active HIGH S-R latch occurs when	Upper Quartile	Lower Quartile	Total Count	%	Disc Index
a) S = 1, R = 0	0	0	0	0%	0.000
b) S = 0, R = 0	0	3	6	14%	-0.273
c) S = 1, R = 1	12	6	33	78%	0.545
d) S = 0, R = 1	0	1	1	2%	-0.091
Marks: (1) Percentage Correct: **78%**	Difficulty Factor: 0.8462				

Figure 33(a) Detail distractor analysis indicating upper- and lower quartiles

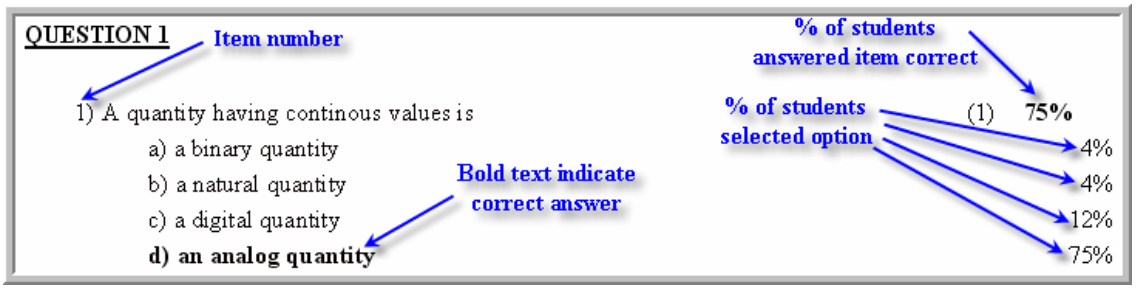

Figure 33(b) Summary distractor analysis

5.3 CONCLUSION

All of these forms of validity need to be addressed in assessment to ensure validity in assessment, and to make sure that the assessment is defensible and, in a sense, fair to students (Maree & Fraser 2004:35). After the item analysis has been completed the question might now be asked: Is item analysis the answer for improving the measurement instrument (test) that we have used? (as discussed in 5.1). The answer is that one cannot have too much confidence in statistical methods of test analysis, as these methods only provide a more objective method to establish whether an accurate estimate of student knowledge has been determined. Item analysis is only another method provided by CCAT in the entire process of determining reliability, validity and fairness. A proposed example of an assessment validity review document is shown in Annexure A that was used by the assessment committee to document the validity of a test.