www.bircu-journal.com/index.php/birci
Budapest International Research and Critics Institute-Journal (BIRCI-Journal)
Humanities and Social Sciences
ISSN 2615-3076 Online)
ISSN 2615-1715 (Print)
Budapest Institute

Psychometric Testing of the Bullying Scale on Students in Indonesia

Uswatun Hasanah[1], Arina Husna Zaini[2]
[1]*Universitas Putra Abadi Langkat, Indonesia*
[2]*UIN Sayyid Ali Rahmatullah Tulungagung, Indonesia*
hasanahuswa23@gmail.com, arinahusna@uinsatu.ac.id

Abstract

This research aims to determine and evaluate bullying measurement tools at the junior high school to high school student level. Researchers considered cases that occurred in Indonesian education with symptoms of verbal, relational and physical violence among students. The number of respondents in this study was around 100 people. The research method used by researchers is incidental sampling. The results obtained in this research are that the alpha reliability value in this research is 0.879. Factor analysis obtained in this research was to test the validity of each dimension based on the results of LISREL software calculations. The norm results stated that the subjects had quite different levels of bullying categories in the percentile rank norms and t-score norms. It can be seen that the distribution of subjects is uneven and shows that some are very low and some are very high. Future research needs to consider subject characteristics to be more specific and detailed.

Keywords
Psychometric; testing; bullying scale

I. Introduction

As time goes by, casesbullyingdeveloping to an alarming level, especially in Indonesian educational institutions. There are several cases of bullying that occur around us, for example the case of bullying at SMA 90 Jakarta. Grade 1 students were forced to take off their clothes, do push ups, run and were slapped. SMA 90 then suspended 31 students involved in bullying for 5 days. The seniors signed a letter of agreement on a stamp so they would not repeat their actions. Seeing that there are many more cases of bullying that occur in the world of education in Indonesia, therefore we conducted research on bullying to find out how bullying occurs and how school children feel who are victims of bullying from seniors and even their own friends.

The definition of bullying according to Olweus (1993) is a long-term and repeated negative action carried out by one or more people against another person, where there is an imbalance of power and the victim does not have the ability to protect himself. Sullivan (2000) defines bullying as aggressive behavior carried out consciously and deliberately by one or more people who are victims with the aim of causing harm. According to Rigby (2003) bullying is a systematic abuse of power in relationships with other people. Duffy (2004) completes the definition of bullying by adding forms of bullying. According to him, bullying can occur in verbal, relational and physical forms.

The bullying measuring instrument comes from Dini (2010) which is an adaptation of a measuring instrument previously created by Sari (2008), because the researcher considers the measuring instrument to be in accordance with what the researcher created. This measuring tool is used to determine participants' involvement in bullying behavior. The importance of re-psychometric testing of the bullying scale is to determine the level

of reliability and validity of the bullying scale. Apart from that, to find out whether this measuring instrument is suitable for reuse to measure the psychological construct, namely bullying.

1.1 Research Problems

Is the scale bullying valid and reliable in measuring bullying in middle and high school students?

1.2 Objective

The purpose of using this test tool is to test the validity and reliability of the bullying scale in measuring bullying in middle and high school students.

1.3 Benefit

a. Theoretical Benefits:

It is hoped that the results of integrating this measuring instrument can contribute to the field of psychology and provide an overview of bullying that occurs in middle and high school students, especially teenagers.

b. Practical Benefits:

By using this bullying scale, we can contribute information to society, especially teenagers, in the form of data on the percentage of teenagers who experience bullying and those who do not experience bullying.

II. Review of Literature

2.1 Definition of Bullying

Bullyingaccording to Olweus (1993) is a negative action over a long and repeated period of time carried out by one or more people against another person, where there is an imbalance of power and the victim does not have the ability to protect himself. Duffy (2004) completes the definition of bullying by adding forms of bullying. According to him, bullying can occur in verbal, relational and physical forms.

2.2 Forms of Bullying

Duffy (2004) completes the definition of bullying by adding forms of bullying. According to him, bullying can occur in three forms, namely:

1. Verbal
   Verbal Bullying related to words or words. Actions included in this type are threatening, mocking, giving inappropriate nicknames, intimidating someone with harsh words, terrorizing over the telephone, threatening, labeling someone, uttering racist, insulting words, and spreading gossip negative.
2. Relational
   Relational Bullying relates to all behavior that manipulates or damages relationships with other people. Actions included in this type of bullying are deliberately silent someone, ignoring someone's whereabouts, isolating someone, slandering and destroying friendships.
3. Physique
   Physical Bullying is the most visible form of bullying because it is direct and there is physical contact between the victim and the perpetrator. Examples of behavior include: restraining, pulling hair, hitting, kicking, pinching, pushing, scratching, spitting, and other forms of physical attacks, including intentionally destroying someone's property.

2.3 Psychometric Theory

a. Test Type

The types of tests are as follows:

1. Based on the type of behavior measured:
   - Maximum performance test (ability test/optimal performance test): An individual's capacity to do something or complete a given task. Because an individual's response is related to his cognitive abilities, the answer given by the individual can be said to be the "right" or "wrong" answer and given an appropriate score (Azwar, 2004). Examples: intelligence tests (WAIS-R, Standford-Binet, etc.), aptitude tests (DAT), learning achievement tests, and learning potential tests (TPA, GRE, etc.).
   - Typical performance test (personality test): A specific set of characteristics and traits that drive the way an individual interacts with individuals or situations. Individual responses have very little to do with cognitive abilities and are typical (typical) of each person, therefore responses in the form of typical performance cannot be said to be "wrong" (Azwar, 2004). Examples: personality tests (Rorschach, Wartegg), attitude scales, and interest inventories.
2. Based on the method of collection or administration:
   - Individual test: given at a time to only one respondent.
   - Group test: can be given to more than one respondent at a time.
3. Based on the nature/form of response:
   - Paper & pencil test (using writing tools).
   - Oral test (verbal-oral).
   - Non-verbal test (using objects/tools)
4. Based on task completion time:
   - Speed Test: a measuring tool whose items are relatively easy, have a short time limit, and place greater emphasis on speed in carrying out tasks.
   - Power test: a measuring instrument whose items have a degree of difficulty, no time limit, and more emphasis on the ability to complete tasks.

2.4 Requirements for Good Measuring Instruments

a. Reliability

1. Understanding Reliability
   One of the requirements for a good measuring instrument is to obtain relatively similar measuring results on the same subject under the same conditions. Reliability is the same as the consistency of the score obtained by someone, when the measurement is carried out again with the same test at different times and with different tests but the items are equivalent (Anastasi & Urbina, 1997).
2. Reliability Coefficient Estimation Method
   In general, there are reliability estimation methods divided into two procedures (Crocker & Algina, 1986):

b. A procedure that requires Two Test Administrations

Procedures that require two test administrations are carried out by administering the test in two takes, to the same subject, with the same test or two equivalent tests.

The procedure that requires two test administrations consists of two methods, namely:

1. Test-retest reliability method
   - Measurement Consistency measured with the same instrument at different times.
   - Reliability coefficient correlation between scores on two measurement results on the same subject (coefficient of stability).

- Used in tests that aim to measure traits or characteristics that remain relatively unchanged over time (temporal stability).
- Error Type time-sampling error.
- Source of Error differences in external conditions & internal such as Maturity, trauma, learning, temperature, experience, noise, counseling/therapy, and instruction.

2. Alternate-form reliability method
   - Alternate-form is the same as Parallel-Form; Equivalent-Form determine the consistency of scores on two equivalent tests.
   - Two tests are said to be parallel if they meet the same specifications regarding: number of items, item form, content coverage, range and degree of item difficulty, instructions, time limit, examples, and format.
   - This method is suitable for tests that are not influenced by the learning process.

c. Procedures That Require One Test Administration/Single Trial/Single Test Administration

Procedures that require one administration are used if it is not possible to carry out a retest, there are no parallel tests, the time given is very limited, and so on.

The procedure that requires one test administration consists of four methods, namely:

1. Split halves

   The test can be divided into two, to see the consistency of the subject's responses in 2 equal halves of the test through internal consistency. To estimate the reliability of all tests, the Spearman-Brown formula was used. The condition is that both hemispheres have a mean & variances are not significantly different. The type of error in split half is content sampling error. Methods used for splitting tests:
   - First hemisphere-second hemisphere The test is divided into two halves, the initial hemisphere and the final hemisphere of the test.
   - Odd-even The test is divided into two parts, odd numbered item parts (odd) and even numbered item parts (even).

   Crocker & Algina (1986) added another method of dividing, namely based on the degree of difficulty & divide randomly.

a) Kuder Richardson

   The KR reliability method is used in tests whose items measure the same trait (homogeneous/item consistency). KR is a formula for calculating the reliability of a test that has dichotomous items (given a score of 1 or 0). The sources of error in the KD reliability method are content sampling and content heterogeneity.

b) Coefficient Alpha

   Coefficient Alpha is the most common method for conducting reliability testing through internal consistency (Kaplan & Saccuzo, 2005).

   Cronbach (1951) created a formula that can be used for non-dichotomous data (score > 1). The purpose of this reliability testing is to see homogeneity/item consistency. If the item is dichotomous, the Cronbach Alpha result is the same as KR20 (both formulas are basically the same). Sources of error in Coefficient Alpha are content sampling and content heterogeneity.

c) Scorer Reliability
   - Reliability is the consistency of test scores obtained by subjects from two or more scorers.
   - Required if the test is open-ended, such as essays, projection tests, observations.
   - The error that occurs is the interscorer difference.

2.5 Interpretation of Reliability Coefficients

Interpretation of reliability must be related to the reliability method used. The reliability coefficient is used to obtain information about the large variability in test scores caused by error of measurement and true score. The reliability coefficient is the proportion of true score variance and observed score variance.

The reliability coefficient limits are as follows:

1) Anastasi & Urbina (1997): generally 0.8.
2) Kaplan & Saccuzzo (1989: regarding test objectives
   - Research: 0.7–0.8.
   - Clinical (diagnosis): 0.95.

2.6 Factors Affecting Reliability

The factors that can influence reliability are as follows:

1. Subject Characteristics.

- Variability (subjects vary in the attribute or trait being measured). Example: restriction of range.
- Ability level (subjects vary in ability).

2. Test characteristics (test length/number of items and representativeness of items).
3. Use of test scores.
4. Reliability estimation method used.
5. Statistical Techniques in Reliability Testing

- Correlation

1) Perform test results scoring.
2) Correlate the results of the first test with the second test.

- Kuder Richardson Cronbach's Alpha

1) Perform test results scoring.
2) Pay attention to the presence/absence of Unfavorable items (if there are, the scoring is reversed).
3) Perform Cronbach's Alpha testing technique.

2.7 Validity

1. Understanding Validity

   There are several definitions of validity, including the following:

- What does the test measure and how precisely does the test measure what it is intended to measure (Anastasi & Urbina, 1997).
- The correspondence between a test score or measurement and the quality that the test is believed to measure (Kaplan & Saccuzzo, 2005).
- Judgment(about tests)which is based on evidence about the accuracy of conclusions drawn from test scores (Cohen & Swerdlik, 2010).

2. Validity Function

   The functions of validity are as follows:

- Explains the attributes/constructs/traits/factors measured by a test.
- Describes how precisely it is measured.
- Explain what can be interpreted/interpreted from a test score.

3. Validation Procedure

   All procedures for establishing test validity must consider the relationship between test scores and other facts (evidence) that are observable and independent of the trait to be measured. The relationship between test scores and criteria is the same as the validity coefficient.

4. Strategy/Approach in Validity Testing

The strategies/approaches in validity testing are as follows:

1) Content Validity/Content-Description Procedures/Content Validity

Content validity is a measure of the extent to which a test is representative for measuring the content of a particular domain of behavior.

2) Criterion-Related Validity/Criterion-Prediction Procedures/Criterion Validity

Criterion Validity is the effectiveness of a test in predicting an individual's performance on a particular activity. There are 2 (two) types of criterion validity, namely:

a. Predictive Validity (forecasting validity)□a measure of the extent to which a test is valid in predicting certain things.

b. Concurrent Validity (diagnostic/concurrent validity)□a measure of the extent to which a test is valid in diagnosing a person's condition or current status in a particular matter.

3) Construct Validity/Construct-Identification Procedures/Construct Validity

Construct Validity isa measure of the extent to which a test measures a certain theoretical construct or trait. Constructs are psychological dimensions that have been formulated clearly, in detail and operationally. A test that is valid for measuring construct X is not necessarily valid for measuring construct Y.

The general procedures for testing construct validity are:

a. Learn the theory surrounding the construct to be measured (for example: intelligence, mechanical compr, verbal fluency, anxiety, and so on).

b. Develop a theory-based hypothesis (link it to a specific technique).

c. Analysis of suitability of empirical test results.

The construct validity testing techniques are as follows:

1) Developmental Changes.
2) Correlation With Other Tests.
3) Factor analysis.
4) Internal Consistency.
5) Convergent and Discriminant Validation
6) Experimental Intervention

2.8 Face Validity

Face validity (validity of appearance/impression) cannot be equated with content validity. Face validity is not validity in the technical sense because it does not measure what the test actually wants to measure, but only the test taker's impression of what the test measures (what the test displays). Face validity functions to build rapport and increase test taker motivation.

5. Interpretation of Validity Coefficients

The interpretation of the validity coefficient (Anastasi & Urbina, 1997) is as follows:

1) It must be significant at a certain level and high enough to be able to identify and distinguish individuals.
2) Related to the purpose of the test.
3) Related to construct theory.
4) Associated with validity testing methods.
5) Specifically for criterion-validity, the correlation is expected to be significant and high.

2.9 Item Analysis

Item analysis is divided into two, namely qualitative analysis and quantitative analysis. Qualitative analysis is also divided into two, namely moderator techniques and panel techniques. The Moderator Technique is a discussion technique in which one person

acts as a mediator, while the Panel Technique is a technique for reviewing question items in which each question item is reviewed based on the rules for writing question items, namely in terms of material, construction, language/culture, correctness of the answer key/scoring guidelines. Carried out by several reviewers. There are two approaches to quantitative analysis, namely the classical approach and the modern approach. The classical approach is divided into three, namely the level of difficulty. This index of level of difficulty is generally expressed in the form of a proportion whose size ranges from 0.00 - 1.00 (Aiken, 1994: 66). The greater the difficulty index obtained from the calculation results, the easier the question is. The differentiating power of a question item can differentiate between test takers who have mastered the material being asked and test takers who have not/less/haven't mastered the material being asked. The benefit of differentiating power is to improve the quality of each question item through empirical data. The higher the item's discriminating power index means that the item in question is more capable of distinguishing participants who have understood the material from participants who have not understood the material. The spread of distractor answers was chosen by at least 2.5%, and more distractors were chosen by the lower group.

2.10 Norm

1. Understanding Norms
   Norms are the distribution of scores from a group that are used as a benchmark to give meaning to individual scores (Broom & Selznic, 1999). In general, there are 2 references that are often used in interpreting scores, including:
   1) Criterion Reference (Criterion-Reference).
   2) Norm Reference.
2. Reference Criteria
   1) Scores that can specifically describe what students know or can do or the material and skills mastered,
   2) Determined by linking it to student achievement on a particular standard or goal.
   3) Can be in the form of a score (pass/fail) or indicate the level of competency mastered.
   4) Suitable for tests that measure basic skills, not complex processes.
   5) Reported in the form of raw scores, usually displayed as a percentage of questions that can be answered correctly. (Note: the questions given must represent each predetermined criterion),
   6) In its use, it is necessary to determine the cut off score. Example: Minimum Completeness Criteria (KKM).

Interpreting the scores obtained by individuals and comparing them with group scores in the same test, which is called the "norm group". Can be used to compare students with other standardized groups. Types of norms:

1. Grade Equivalent Scores, Student scores are interpreted according to the class average ability. The reference is the average score obtained by the standard sample in a particular lesson. This is one of the scores that often causes misunderstandings.
2. Age Equivalent Score, the reference is the average score obtained by the standard sample age group on the test. Often misinterpreted.
3. Percentile Ranks, shows a person's relative position in a standardization sample.Considered as a sequence in the position of a group consisting of 100 people where better is the subject with a greater percentile value. The percentile score is the percentage of the number of people who are below the raw score (a certain raw score), Percentile ≠ Percentage.

How to calculate the Percentile Rank of each Raw Score:

$$\frac{Cfi + 0.5(fi)}{N} x100$$

Information:

Cfi : cumulative frequency lower limit of real score
Fi : score frequency
N : number of subjects in the sample

4. Standard Scale Scores,scores resulting from transformation into a normal distribution. Overcoming Percentile Rank limitations in terms of using the same units. Can be used to compare individual abilities on two different tests. Z-score, a score that shows how far a student's ability is compared to the class average in standard deviation units. Score Interpretation: positive/negative sign indicates that the student's score is higher than the class average (+) or lower than the class average (-). The z-score value shows the distance between the student's score and the class average, in standard deviation units. The standard score reflects a normal distribution form.

III. Research Methods

3.1 Research Subject

The population in this study were high school students, and we took samples from high school students who were in public places such as schools, malls and others with an age range of 14 - 17 years, male and female. The sample size was 100 people. The sampling technique in this research is incidental sampling.

3.2 Measuring Instrument

The measuring instrument used to obtain the data needed in this research is the bullying scale. Bullying is a long-term and repeated negative action carried out by one or more people against another person, where there is an imbalance of power and the victim does not have the ability to protect himself.

3.3 Classification of Measuring Instruments

The classification of measuring instruments in bullying tests is classified based on:

1. The type of behavior being measured□typical performance test, because there is no right or wrong in the answer choices given, the answer choices are given to measure the subject's personality.
2. Method of collection/administration□group, because at one time the measuring instrument can be given to more than 1 person.
3. Nature/form of response□paper and pencil test, the research was carried out using writing tools that had been prepared by the researcher.
4. Test completion deadline□speed test, the time in this study is short and only looks at the final results of the measuring instrument used by the subject.

3.4 Item Grid

- Construct: Bullying

□ Verbal

- Operational definition: actions included in this form are threatening, mocking, harassing, giving inappropriate nicknames, intimidating someone with harsh words.

- Indicators: being teased by certain students when passing in front of them; and being called “special” names that embarrassed me.

 Relational
- Operational definition: actions included in this form are deliberately silent someone, ignoring someone's existence, isolating, causing negative gossip, slandering.
- Indicators: silence by a group of students when carrying out tasks in a discussion group; and shunned by a certain group of students.

 Physique
- Operational definition: actions included in the form are pushing, hitting, kicking, punching, scratching, pulling, tearing clothes, destroying books, destroying/stealing other people's belongings.
- Indicators: found that my belongings were deliberately damaged by certain students; and was intentionally hit by another student violently.

3.5 Scoring Techniques

For each item there are 6 categories of answer choices, namely Never (TP), Almost Never (HTP), Rarely (JR), Sometimes (KD), Often (SR), and Always (SLL). The rating system on the bullying scale for the favorable type is 1 for the answer Always (SLL), 2 for the answer Often (SR), 3 for the answer Sometimes (KD), 4 for the answer Rarely (JR), 5 for the answer Almost Never Never (HTP), and 6 for the answer Never (TP). On the other hand, the unfavorable type is 6 for Always (SLL), 5 for Often (SR), 4 for Sometimes (KD), 3 for Rarely (JR), 2 for Almost Never (HTP), and 2 for Almost Never (HTP). 1 for the answer Never (TP).

Measuring Instrument Testing Procedures

1. Preparation phase
- Determine the psychological constructs you want to use in research.
- Look for measuring instruments according to the dimensions of the construct you want to use in research.
- Instruments in the form of questionnaires were distributed to research respondents.
- Then data processing was carried out using SPSS.
- After that, reliability, validity, item analysis and norm testing is carried out using an appropriate program.
- Data analysis from the tests that have been carried out is carried out.
- Lastly, make a research report.

2. Data Collection Stage
- The data collection locations in this research were schools, malls and other public places.
- Data collection was carried out on 6–10 November 2014.
- The implementation in this study was individual and there were no knocks.

3.6 Psychometric Techniques Used

- Reliability Testing Techniques
  The reliability testing technique used in this research is Cronbach's alpha. This is because the measuring instrument used uses a Likert scale. Apart from that, the data contained in the instrument is non-dichotomous or polytomous.
- Validity Testing Techniques
  The validity testing technique used in this research is construct validity. This is because construct validity can measure a certain theoretical construct or trait that has been formulated clearly, in detail and operationally based on the LISREL program.

- Item Analysis Techniques
  The item analysis technique used in this research is internal consistency, to see the significance of each item and see its differentiating power.
- Norm Formulation Techniques
  The norming technique used in this research is within group norm, looking at norms based on groups that fill out the bullying scale.

IV. Result and Discussion

4.1 Implementation of Data Collection

This research was conducted in schools, malls and other public places on 6–10 November 2022. The subjects in this research were middle school and high school students aged 14–17 years, with a sample size of 100 people.

4.2 Reliability Testing Results

The alpha reliability value in this study was 0.879. Positive instrument test results and > 0.8 is said to be reliable. Because the alpha value test result is 0.879, the instrument is reliable. Item results < 0.879 is a valid item, otherwise each item > 0.879 is a dropped item. So, none of the 14 items in the bullying scale were dropped because the 14 items were valid.

SEM value (*Standard Error of Measurement*):

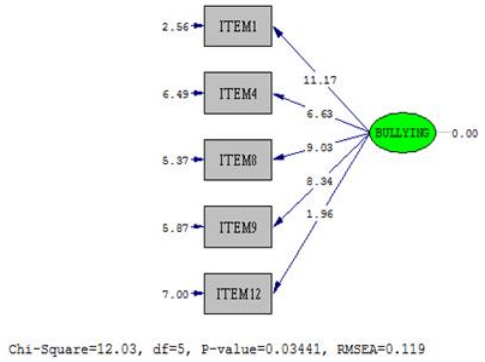
$$SEM = SD\sqrt{1 - r_{xx}{}^2}$$

SEM = standard error of measurement
SD = standard deviation of test scores
r_{xx} = reliability coefficient

SEM value (*Standard Error of Measurement*) in dimension 1 it is 0.496, dimension 2 is 0.484, and dimension 3 is 0.555. The expected standard error is relatively small, namely 0.5 or 0.4. This is in accordance with the standard error values obtained from these three dimensions, so it can be said that the standard error on this bullying scale is low.

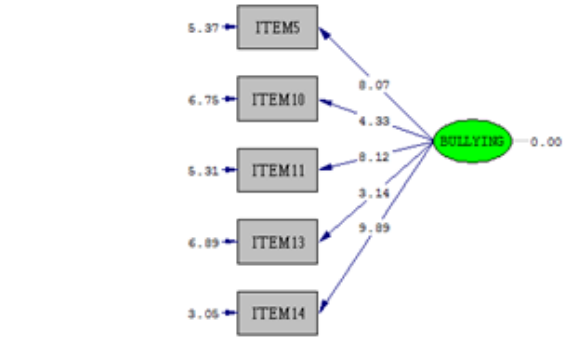
4.3 Validity Testing Results

Construct Validity – Factor Analysis:



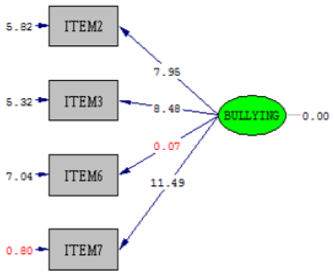
Chi-Square=12.03, df=5, P-value=0.03441, RMSEA=0.119

In dimension 1, namely verbal, all items (item 1, item 4, item 8, item 9, item 12) are valid and reliable. The RMSEA value of dimension 1 is 0.119 > 0.08 which means invalid, the chi-square value is 12.03 > 0.001 means invalid and the P-value = 0.03441 > 0.05.



Chi-Square=20.29, df=5, P-value=0.00110, RMSEA=0.176

In dimension 2, namely relational, all items (item 5, item 10, item 11, item 13, item 14) are valid and reliable. The RMSEA value of dimension 2 is 0.176 > 0.08 which means invalid, the chi-square value is 20.29 > 0.001 means invalid and the P-value = 0.00110 > 0.05.



Chi-Square=11.78, df=2, P-value=0.00277, RMSEA=0.222

In dimension 3, namely physical, item 6 is not valid but reliable, item 7 is valid but not reliable, while item 2 and item 3 are valid and reliable. The RMSEA value of dimension 3 is 0.222 > 0.001 means invalid, and P-value = 0.00277 > 0.05.

4.4 Item Analysis Results

DIMENSION 1

In conducting item analysis we use internal consistency to seeThe significance of dimension 1 which consists of items 1, 4, 8, 9 and 12. All of these items have a significance of 0.000 with the information that the item is valid and has distinguishing power that the question is accepted.

DIMENSION 2

In conducting item analysis we use internal consistency to seeThe significance of dimension 2 which consists of items 5, 10, 11, 13, and 14. All of these items have a significance of 0.000 with the information that the item is valid and has distinguishing power that the question is accepted.

DIMENSION 3

In conducting item analysis we use internal consistency to seeThe significance of dimension 3 which consists of items 2, 3, 6, and 7. All of these items have a significance of 0.000, while items 5, 7, 15, 21 and 24 have a significance of 0.000 with the information that the items are valid and have differentiating power that the matter was accepted.

4.5 Results of Norm Preparation

There are 20 people who are in the percentile rank ≤20 and have a total score of ≤25, which is in the very low category, there are 21 people who are in the percentile rank between 21-40 and have a total score of 26-31, which is in the low category, there are 22 people who are in the percentile rank 41–60 with a total score of 32–37, including in the medium category, there are 18 people who are in the percentile rank between 61–80 with a total score of 38–50, including in the high category, and there are 19 people who are in the percentile rank >80 with a total score of >50 is included in the very high category.

There are 20 people who have a t-score ≤40 with a total score of ≤25, which is in the very low category, there are 42 people who have a t-score between 41–50 and have a total score of 26–37, which is in the low category, there are 15 people who are on a t-score between 51–60 and have a total score of 38–49 are included in the medium category, there are 22 people who are on a t-score between 61–70 and have a total score of 50–61 are included in the high category, and there are 1 person who has a t-score >70 with a total score >61 is included in the very high category.

In the t-score norm table, there are 2 people who have very low and very high scores, namely 1 person has a t-score ≤40, namely 31.09 with a total score of 14 which is in the very low category, while for 1 person who has t-score >70 is 71.63 with a total score of 62 which is included in the very high category. From these results it was found that there were differences in subjects who had different abilities.

V. Conclusion

Based on the results of reliability, validity, item analysis and norm testing carried out by researchers, it was found that the results of reliability testing for the instrument were reliable. The results for validity testing show that there is one item from the three dimensions that is not valid but reliable, namely item 6, but there is also one item that is valid but not reliable, namely item 7, and there are several other items that are both valid and reliable. For analysis of items from each dimension, the items contained in the three dimensions have a significance value of < 0.05 and has valid information with the distinguishing category "Question Accepted". Meanwhile, for the norm results, it was found that the subjects had quite different levels of bullying categories in the percentile rank norm table and the t-score norm table. It can be seen that the distribution of subjects is uneven, some are very low and some are very high.

This research has a fairly good reliability coefficient. The items in this research have achieved the target. Apart from that, the subjects in this research are representative. The norms in the group resulting from this research should show a high bullying value, but in this study it shows a low bullying value. From the norm table that has been created for both the percentile rank norm table and the t-score norm table, the subjects have different bullying scores, some are very low, some are medium to very high. This can be seen from the norm table contained in the journal.

For other researchers who are interested in using this measuring instrument, it is best to pay attention to the items to be studied, such as: researchers must be able to adjust the characteristics of the subject to each individual who will be researched. If you are going to research middle and high school students, the characteristics of the subject are more specific. and adjusted to the items in the measuring instrument. As well as paying attention to theories that can clarify and sharpen the aspects that will be used so that the desired goals can be achieved.

References

Anastasi, A. & Urbina, S. (1997). *Psychological Testing 7th edition*. New Jersey: Prentice-Hall Inc.

Anastasia, A. & Urbina, S. (2007). *Tes psikologi, edisi ke tujuh. Jakarta*: PT. Indeks.

Atwater, E. & Duffy, K. G. (2004). *Psychology For Living: Adjustment, Growth and Behavior Today 8th edition*. New Jersey: Prentice-Hall, Inc.

Azwar, S., (2004). *Dasar-dasar psikometri*. Yogyakarta: Pustaka Pelajar.

Azwar, S., (2012). *Reliabilitas dan validitas (Ed. 4).* Yogyakarta: Pustaka Pelajar.

Broom, L., & Selznick, P. (1955). *Sociology 11th edition*. New York: Harper International.

Cohen, R. & Swerdlik, M. E. (2010). *Psychology Testing and Assessment*. Mc Graw Hill Higher Education: United State.

Crocker, L. & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. New York: Harcourt.

Dini, (2010). *Hubungan antara bullying dengan school well-being pada siswa SMA*. Skripsi. Depok: Fak. Psikologi UI.

Harrison, S. & Sullivan, P.H. (2000). *Profiting from Intellectual Capital: Learning from Leading Companies*. Journal of Intellectual Capital, 1, 36-44.

Kaplan, R.M & Saccuzzo, D.P. (2012). Pengukuran Psikologi: Prinsip, Penerapan, dan Isu. Edisi 7. Jakarta: Penerbit Salemba Humanika.

Kaplan, R.M. & Saccuzzo, D.P. (2005). *Psychological Testing Principles, Application and Issue*. Sixth Edition. USA: Wadsworth.

Olweus, D. (1993). *Bullying at school: What we know and what we can do*. Malden, MA: Blackwell Publishing.

Rigby, K. (2003). *Consequences of bullying in schools*. The Canadian Journal of Psychiatry, (48) 9, 583-590.

Sari. (2008). *Sistem Pakar dan Pengembangannya*. Yogyakarta: Graha Ilmu.