Paola Molina & Ana Muntean

1 Evaluation tools: notes on definition, reliability, validity and administration

In this book evaluation tools on different aspects of play are presented (see Chapter 3, Bulgarelli, Bianquin, Caprino, Molina & Ray-Kaeser, 2018). A preliminary consideration about these tools concerns their validity and reliability, aspects that allow to consider them as *tests* and their ethical use.

But what are tests indeed? What features do they need to have in order to trust on them? Why are they useful?

1.1 History of development of test

A short historical overview of these instruments may be useful (Gregory, 2014). The first examples of the tools that can be compared to the modern tests date back to antiquity: in the China empire, particular evaluation procedures were intended for selection of Mandarins (staff selection); in ancient Greece, various philosophical schools used specific tools to evaluate pupil's learning (profit tests). The first true tests, however, were born between the late nineteenth and the early twentieth centuries, during and in relation to the development of psychology as a science:

- In the psychophysical laboratories in which experimental psychology was born (Wundt).
- For clinical purposes, i.e. to differentiate people suffering from intellectual disability from mentally ill individuals, who show the same intellectual performance but because they suffer from psychic problems: these instruments were developedand utilized by the psychiatric scholars as Esquirol or Séguin.
- To study individual differences, the Galton's goal. Galton himself, and then his students, among them the most well-known is Cattel, develop a series of tests, especially of physiological and sensory type (ranging from the size of the skull to the force of the handshake or the sensory thresholds), which in their intentions should provide an assessment of the intelligence of individuals: unfortunately, the scores obtained with these tests showed no relation to success in life or in academic path, success that should be associated with high intelligence. Nevertheless, the evaluation method remains an intake consolidated for a subsequent research.
- To answer practical problems given by the extension of compulsory education: the first true intelligence test was published in 1905 by Binet and Simon. They were instructed by the Ministry of Education to design a screening of children whose intellectual level did not allow to benefit from normal school, and subsequently to include them in special instruction classes. Unlike the previous

authors, Binet and Simon thought that intelligence could be better measured by means of the higher psychological processes rather than the elementary sensory processes such as reaction time. In the Scale published in 1905 (Binet and Simon, 1905), the items were arranged by approximate level of difficulty instead of content, level established by the examination of typical responses of children from 3 to 11 years: a sort of rough standardization (see later, p. 12). In accordance with the definition of tests, the Binet–Simon Scale is considered as being the first true test.

Over the following years, the design and the use of tests were widespread: tests that measure intelligence have been widely used by the US Army for the selection of soldiers and officers. For the same purposes, the first aptitude test batteries have been developed, which measure specific skills tailored to specific tasks (for example, the visual acuity or the reflexes required for a pilot). In education as well, standardized tests have partially replaced the oral examinations, that are more time-consuming and considered less objective and more subject to individual distortion. In the ’70s of the last century, a crisis hit the use of tests, mainly because of the indiscriminate use that had been made, with little control over the quality of the instruments and their administration: the tests were considered unfair, especially towards ethnic and linguistic minorities. In the United States, where the criticism movement was born, the result was a more rigorous methodology, coupled with a greater prudence in use, particularly in the field of education: for instance, the American Association of Psychology (APA) proposed the *Standards for Educational and Psychological Testing* (APA, 1992), which have become a worldwide reference point for the educational and psychological tests. There is currently a recovery in the use of tests, which are more rigorous from the methodological point of view and are applied with greater awareness.

1.2 Definition of tests

Tests are tools that psychologists and other professionals use in order to collect data about people (Groth-Marnat, 2003). Considering the ever-increasing plethora of online assessments, this definition, based on the professional interaction between the psychologist and the client, seems less than adequate. A test is an instrument which asks test-takers to perform some measurable or observable behaviour, the intention being to highlight personal characteristics which are not particularly evident, but nevertheless, salient for providing an understanding of the person and the predictability of their behaviours. Tests are considered to be one of the greatest achievements in psychology and are used for the assessment of human behaviour throughout all areas of human activity, examples being health care, education, justice, social protection, industry and transport, and entertainment.

A huge range of behaviour can be explore through different types of tests, including cognitive levels and achievements, human development and personal behaviours, personality and psychopathology, skills including driving safety and academic/educational aptitudes, neuropsychological, language and sensory-motor aspects, and social and vocational characteristics.

Applying tests in assessment provides a great deal of information in a short time and can highlight characteristics of which the subject being assessed is sometimes not aware. The use of tests is based on a number of important assumptions such as respondents' truthfulness and accuracy in their answers and awareness of the risks of the occurrence of errors due to the instrument itself, the respondent, the examiner or the environmental conditions.

Tests are specially designed to highlight individual outcomes for children or adults. For this reason, tests are used for psychological assessments within clinical work or in research as measurement tools intended to prove or correct the hypotheses of a clinician or researcher, to foster predictability and to orientate interventions.

The choice of a specific test is based on the theoretical foundation of the test, its psychometrics characteristics (standardization, reliability, validity) and practical considerations regarding the administration procedure (Groth-Marnat, 2003). Differences exist between tests as assessment instruments in terms of their field of investigation and the goal of the evaluation, the method, the time required for test administration, the content, structure and theoretical orientation, the performed behaviour elicited and the sample of behaviours they are intended to measure, the procedure for scoring and interpreting the results, and above all their psychometric characteristics.

1.3 Test characteristics

The tests, therefore, havea long history, but why such tools are important?

The main function of tests is to allow an evaluation free from subjective bias present in the everyday life. In fact, people's judgments are influenced by a number of factors (partly aware, partly unconscious) that do not always allow them to be objective. For example, people are influenced by the characteristics of the stimuli: more frequent facts (for example, the usual delay), or intense, or exceptional facts (a very intelligent or very stupid answer) are more easily impressed in our memory, and therefore weigh more on our judgment; the information gathered as the first or the last remain longer in mind, etc. Moreover, in evaluating others' stereotypes, implicit personality theories and expectations, the perceived attribution of features on the basis of difference/resemblance with the evaluator, etc., play an important role.

The tests are useful because they grant, as much as possible, an evaluation free from subjective bias. To be a test, a tool shall offer a series of guarantees about what it measures and how it can do so: a test consists essentially of an objective

and standardised measure of a sample of behaviours (Anastasi, 1968). Mainly, a test is a set of verbal or non-verbal tasks, called *items*, proposed to the subject. The set of items is a representative sample of behaviours, directly observable, in which the *competence* measured by the test is revealed. This competence, called *construct*, is instead a psychic quality, not directly observable, which is translated (*operationalized*) through observable behaviours that are evaluated by the test items. For example, I cannot directly observe *aggressiveness* or *anxiety* (constructs), but I can ask the subject if he or she is reacting with a threat when someone unfairly overtakes him or her on the highway, or if he or she bites his/her nails in the waiting room of the dentist. These responses, the items in the test, can be considered as indicators of aggressiveness or anxiety, and a sufficient number of items can discriminate people along a continuum that goes from the *low* presence to the *high* presence of the construct (aggressiveness or anxiety).

Obviously, for a test to work properly, a series of requirements have to be present in order to ensure *objectivity*, relevance to the construct to be measured (*validity*), and accuracy of the measurement (*reliability*).

The first aspect to be considered is the *uniformity of the administration and scoring* procedures: the first significance of test *standardisation* refers to the the administration of the test. The examiner has to give all the instructions in the same standardized way following the test protocol. The model of the test procedure is the model of experimental research. All subjects are observed in equal conditions, and their performance is evaluated in the same way: differences in response among subjects are therefore determined not by differences in the test to which they are subjected, but by true individual differences in the construct measured by the test. Standardization of the procedure requires careful monitoring of the material used, the instructions, the conditions of administration, etc.: the environment in which the test is administered, or even the moment in the day, may have more or less important effects on the performance of individuals (Bronfenbrenner, 1977). In particular, where the administration is individual and a true relationship is established between the administrator of the test and the person to whom it is administered, the administrator must have adequate preparation both in interaction management in general and with regard to the specific instrument. The score calculation must also give the same guarantees of invariance with respect to the different administrators, so the procedure must be defined in a comprehensive and unambiguous manner.

The test must also give assurances that it can measure what it actually states to measure, i.e. its validity: in fact, the constructor's subjective conviction that the test items properly translate the construct is not enough, but whoever builds the test should provide evidence of this link. The validity starts with and is based on the clear purpose of the test. *Face validity*, that is, the fact that items are convincing for those who submit or use the test, is only the first step of validation. The theory that has allowed the test to be made has to be explicit, and the test should prove to be a good translation (operationalization) of this theory. First of all, tests must be

a comprehensive and adequate sample of the competence they intend to evaluate (*content validity*). In addition, if theory hypothesizes, for example, that males and females have different spatial orientation capabilities, then the builder of the test that evaluates this competence, will have to report research data showing that males and females actually get different scores (*construct validity*). Moreover, evidence of the possibility to predict the performance of subjects in related fields based on test scores must be demonstrated (*predictive validity*): for instance, a good score on entrance test at the university should be able to predict student outcomes in terms of success in obtaining the graduation.

Finally, the accuracy and stability with which the test score measures the construct must be indicated. This feature is called test reliability, and the proof of reliability must be provided by the researcher:

- evidence of the test functioning *stability* over time: if the conditions remain unchanged, a subject should receive the same score in two subsequent test sessions (*Test-retest reliability*);
- evidence of the *independence* of the score from the specific item choice (a relationship must be present among different selections of the items);
- evidence of the proximity between the score obtained by the subject in that particular administration and its true competence (*Scorer and Inter-scorer reliability*), although a measurement error is unavoidable: the better the test, the lower the *confidence interval*, the distance of the score obtained in one single trial, influenced by random factors that can intervene both in raising and lowering the performance, from the *true score* of the subject.

When the coder's judgment is relevant to the scores, as in the projective tests, it is important that the scoring instructions are clear and unique, so that several administrators evaluate the performance equally: the researcher must also provide the value of the *agreement* between different judges evaluating the same test of a subject.

However, the most important effort in test building is the collection of an adequate *standardisation sample* (the second use of the term "standardisation"). In fact, the score obtained by the individual in a test (*rough score*) is not entirely informative. For example, the number of items passed by a 6-year-old child in a cognitive development test do not allow to understand whether his/her performance is better or worse than the standard for children of his/her age. To know this, a large number of children have to be tested (*standardization sample*) and the average performance of children of a given age have to be calculated: then, the score of a particular 6-year-old child can be compared to the performance of this sample, which is the test *norm*, and a new score (*standardized score*) is attributed to the child, score that will put his/her performance across standardization performances. In this way, it is possible to observe if the score of that particular 6-year-old child is average, above or below the average for his age.

1.4 Ethical considerations concerning assessment and child's play assessment

In 1953, the American Psychological Association (APA, 2002) laid down ethical principles requiring that all psychologists should perform services, such as assessment and psychotherapy, which are in the best interests of their clients. Since 1953 several revisions of the first document have been made, the most recent being in 2010, in order to protect and guarantee the human rights of test-takers. The APA is the professional body responsible for rules and regulations in the field of psychology, whether practice or research. In accordance with APA principles, every country develops specific ethical rules for the use of assessment instruments for the benefit of people and to avoid any harm being caused to or misconduct practiced towards them. Criticism of assessment has focused on aspects such as confidentiality, invasion of privacy, cultural bias, and the use of tests that were inadequately validated or used within inappropriate contexts (Groth-Marnat, 2003). The APA ethical principles require the professional doing psychological assessment for psycho-diagnostic or for research purposes to be aware not only of the psychometric adequacy of a test but also of the appropriateness of its use and the potential psycho-social consequences of applying such tests (Messick, 1979). Although some aspects remain controversial, ethical standards are in place and apply to all phases of assessment, starting with the reasons for carrying out the evaluation, the choice of tests to be used, the storing and interpretation of data, and the communication and use of results. These standards have to do with the professional's relationship with the client, a relationship which should contribute to the accuracy of test results and must in no case be harmful for the client. Stringent rules prevent the invasion of the client's privacy, chiefly by requiring the professional to provide clear explanations regarding test and testing relevance, and once this has been done to ask the consent of the client. Conscious that the results of the psychological testing do not provide the degree of reliability expected from laboratory analysis, the examiner should avoid labelling and should not give rigid psycho-diagnoses of the behaviour of the test-taker. Due to the developmental process this requirement is of particular relevance when doing assessments with children. All the above warnings are connected with a competent use of tests. In order to use a specific test, the examiner must have specific training and regularly update their knowledge concerning the use of that test. Psychologists' area of responsibility for psychological testings also covers the vigilance they are required to exercise in order to prevent any use of such tests by unqualified persons. Accuracy in interpretation and the ethical use of test results are also provided for by the competence of well-trained examiners. This professional competence must be matched by appropriate skills for communicating test results. During this final phase of the assessment process, the professional should take care when selecting the receiver of the information and the language to be used with the client, bearing in mind their level of education, their familiarity or otherwise with the test and assessment and especially any possible

emotional reaction on the part of the client. The code of conduct for psychologists provides clear rules for the maintenance of test security by requiring the tests to be kept locked away in a secure place and preventing untrained person gaining access to test materials. Other aspects of security and limitations on the use of assessment results are stipulated within the ethical guidelines for dealing with psychological tests.[1]

Assessment of children whether carried out within a clinic or for research purposes necessitate additional rules designed to promote the best interest of the child. The ethics of child assessment is a complex issue involving: consent of the parent or other legal representative, the consent and especially the assent of the child, and an ethical attitude and adequate knowledge on the part of the professionals. The Convention on the Rights of the Children (CRC) is enshrine international law lays down principles and take account of the child's individuality and dignity, and promote the human rights of children. According to the UNCRC the child has the right to express an opinion and to be taken seriously by adults. The Convention recognizes the children's right to take decisions about important aspects of their life in accordance with their capacities, the cultural context, their life experiences and the support available (see Roth et al., 2013). Reflecting CRC principles, countries around the world have developed their specific national regulations for child assessments. In some countries, ethical regulations and standards laid down by ethical bodies lead to slightly different approaches in regards to child assessment and to the informed consent required from parents and child. The parental consent requirement is based on the parent's duty to protect the child from any possible harm or manipulation and on the child's lack of capacity to take responsible decisions. However due to possible conflicts of interests (particularly when child's assessment is carried out in the context of violence against children perpetrated by a parent) the parent may withhold consent. For situations when child's health is being put at risk specific regulations for waiving parent consent are set-up (CIOMS, WHO, 2016). Situations of child disability can make more critical the need for a waiving of parental consent for the child to be assessed. A further issue dealt with in different ways by some national regulations concerns the relationship between the child's age and capacities and the giving of informed consent.

Child assessment for research purposes involves some specific considerations depending on the domain being investigated: health, psychology, sociology, or social work. The standard recommendation is that whenever possible adult subjects rather than children should be involved. Confidentiality and non-discrimination are important issues in any evaluation that uses child subjects. The limits of confidentiality are connected with the responsibility of professional to promote the best interest of the child. For this reason, the researcher cannot assume unconditional confidentiality when requesting informed consent from parents or from children

1 For further information please visit: http://www.apa.org/ethics/code/

participants in research. Depending on the child's capacities the researcher may ask for the informed consent of the child. However, there are legal restrictions on the use of a child's informed consent. These limitations are based on the child's capacity to fully understand the consequences of taking part in the research. Therefore, the child's continuing assent (agreement) to participation is vital and must be taken into consideration by the researcher even when the child has given their consent. If the child shows unwillingness to be part of the research or to continue the process of assessment, this will put a stop to any further assessment despite the parent and child having initially given their consent. Children assent (agreement) is necessary not only to maximize the accuracy of data collected but also in order to safeguard an important research principle: to avoid causing any harm to the child and to carry out the assessment only in the best interest of the child. In the field of health research, the Council of International Organizations of Medical Sciences (CIOMS) in cooperation with the World Health Organization (WHO) has prepared a set of the "International Ethical Guidelines for Health Related Research Involving Humans" which include special provision for children and adolescents and stipulate *Specific protections to safeguard children's rights and welfare in research* (CIOMS, WHO, 2016, p. 65). This document also discusses the discretionary waiving of the requirement for parental consent on the basis of the principle of assuring the best interests of the child.

The paramount characteristic of the child is playfulness which becomes visible and accessible to assessment through the child's play. Most tests used to assess children in clinic clinical situations and in research are focused on the child's play. Play is the basic language the child uses to communicate about their present situation, their previous experiences and their knowledge concerning the world. Assessment of play is carried out in order to provide an understanding of developmental issues and of the impact of experiences to which the child has been exposed. Play assessment whether carried out using a range of specific tests or based on the observation of a child's spontaneous play forms part of clinical work with children. Some ethical considerations for play assessment are different depending on whether clinical work or research is in view. Even in the case of children of 12 or 13 who possess cognitive maturity, the parent's consent has to be given in order for the child to be assessed. Usually when child's assessment takes place within the clinic, the parent consent is implicit. Child assent is the most important restriction on assessment. Play is genuine and spontaneous and therefore when the assessment is carried out through play the child's assent to assessment is implicit. Ethical considerations require child assessment to be carried out in the same places in which support and help for the child are available and the clinical seitting fulfils this criterion.

The issue of ethical requirements in relation to child assessment is a new topic. With regard to the specific subject of assessment of the child's play and playfulness, there are as yet only very few comments and suggestions in the professional literature, and these are based more on the specific features of childhood than on the human rights context.

1.5 Conclusion: Some considerations regarding the evaluation tools for the play and playfulness

Following the previous indication, it is clear that building a reliable and valid test is not easy: it takes years of work, and a constant subsequent validation work, with the help of the different researchers who use it. Many of the tools that are called *test* in the everyday language only share one or few of these features; sometimes they do not share any of them. Then, what caution should we use? First of all, one should keep in mind that not everything that is called test is really a test: if there is no evidence of standardization, reliability, validity, presence of an adequate normative sample, that is not a real test! Such instruments must be used cautiously, because they are not granted by the procedure necessary to build a test.

Nevertheless, for the specific use and conditions, it is important to have other tools, even if not so robust: this is the case of the evaluation of play and playfulness. In fact, the tools utilized for this type of evaluation are principally built for research, clinical or educational purposes, however, some of the aspects relevant for the test are not relevant for these tools. The forms provided for each instrument in Chapter 3 (Bugarelli et al., 2018) present data on validity, reliability and standardization of the considered instruments, and this information must be attentively considered to choose the more useful and reliable tools.

Perhaps the most important aspect for this type of tools, which are mainly based on observation, is the interrater agreement, which guarantees the possibility to use the evaluation tool consistently: therefore, it is necessary not only to pay attention to the evidence of agreement furnished by the authors of the instrument but also check the agreement in the certain/actual use of the tools.

Another aspect to be considered is the cultural difference in test responses: it is very difficult to obtain a standardized sample for each culture or each language, and frequently the only possibility is the use of a tool standardized for another context. In this case as well, it is necessary to be cautious about possible differences linked to the original cultural context in which the tool was developed.

Finally, an important aspect to be considered is the adequacy of the tool in respect to the specific impairment of the children to observe. Often, an adaptation of the test is possible, although in these cases it is difficult or even impossible to obtain a real validation of the tool. In other cases, different tools are available, but they cannot be suitable for every type of difficulties: for instance, the Vineland Adaptive Behaviour Scales (Sparrow, Balla, & Cicchetti, 2005) are considered a good substitute of typical IQ scales to evaluate the intelligence of disabled children in an everyday context; nevertheless, the VABS are reliable for children with intellectual disability or autism, but are not sufficient for children with severe motor impairment.

References

Anastasi, A. (1968). *Psychologicaltesting* (3rd ed.). Oxford (UK): Macmillan.

American Psychological Association. (2002). Ethical principles of psychologists and code of conduct. *American Psychologist*, 57, 1060-1073

Binet, A., & Simon, T. (1905). *La mesure du développement de l'intelligence chez les jeunes enfants*. Paris (F): Armand Colin.

Bulgarelli, D., Bianquin, N., Caprino, F., Molina, P.,& Ray-Kaeser, S. (2018). Review of the tools for play and play-based assessment. In S. Besio, D. Bulgarelli & V. Stancheva-Popkostadinova (Eds.), *Evaluation of Children's Play. Tools and Methods(pp. 58-113)*. Warsaw (P): De Gruyter Poland.

Bronfenbrenner, U. (1977). Toward an experimental ecology of human development. *American Psychologist, 32*(7), 513-531.

Gregory, R. J. (2014). *Psychological testing: History, Principles and Applications* (7th ed.). London (UK): Pearson.

Groth-Marnat, G. (2003). *Handbook od Psychological Assessment* (4th ed.). Hoboken (NJ): John Wiley &Sons.

Messick, S. (1979). *Test validity and the Ethics of Assessment: Research Report.* Princeton (NJ): Educational Testing Service.

Roth, M., Voicu, C., David-Kacso, A., Antal, I., Muntean, A., Bumbulut, S., & Baciu, C. (2013). Asking for Parental Consent in Research on Exposure of Children to Violence, *Revista de Cercetare si Interventie Sociala, 42*, 85-100.

Sparrow, S. S., Balla, D. A., & Cicchetti, D. V. (2005). *Vineland Adaptive Behavior Scales - Vineland II* (2nd ed.). San Antonio (TX): Pearson.