ASSESSMENT IN CRITICAL THINKING:

A PROPOSAL FOR DIFFERENTIATING WAYS OF THINKING

Carlos Saiz* y Silvia F. Rivas**

***University of Salamanca and ** IE University, Segovia. Spain.**

Article Published in:

***Ergo, Nueva Época* (2008), 22-23, 25-66.**

Contact:

Dr. Carlos Saiz
Universidad de Salamanca
Facultad de Psicología
Avda. de la Merced, 109-131
37005 Salamanca. Spain
☎ +34 923 29 45 00. Ext. 3278
🗐 +34 923 29 46 08
✉ csaiz@usal.es
💻http://www.pensamiento-critico.com/pensacono/autor.htm

Dr. Silvia F. Rivas
IE Universidad Segovia
Facultad de Psicología
Campus de Santa Cruz la Real
C/. Cardenal Zúñiga, 12
40003 Segovia. Spain
☎ +34 921 412 410.
✉ silvia.fernandez@ie.edu
💻 http://www.ie.edu/universidad/es/

Summary.

In the present work we study the different ways currently available for assessing critical thinking. The interest in measuring intellectual competencies stems from the need to check the efficacy of the intervention initiatives applied. It is important to know whether an instruction program actually manages to improve the proposed thinking skills However, the assessment of critical thinking faces a prior hindrance; namely, delimiting what is understood by the term. There are many ways of defining what critical thinking is and, depending on what we believe it might be, we shall act accordingly in its assessment. The way of evaluating intellectual skills to a large extent depends on what we understand by them. Most of the different attempts at evaluation developed to date have been shown to be plagued by severe problems of validity, which has cast doubt on their viability. The aim of the present work is precisely to propose a way of assessing critical thinking that will solve the most important problems inherent to most such initiatives. The work is organised thus: first, we address the importance of the assessment of critical thinking and then we explore the problems present in the different ways of assessing it and consider the viability of a measurement of this nature. Finally, we report the details of our evaluation proposal.

THE NEED TO ASSESS GOOD THINKING

Why do we need to assess our intellectual capacity? Is it to evaluate our aptitudes? When addressing these skills, it would appear that the greatest interest is always sparked in the field of education, and an important aim of education is to evaluate its own results. Regarding the development of thinking skills, it is also necessary to measure whether a given manner of instruction works or not. The benefit in this would be in seeing whether people's performance improves after receiving instruction in this line as compared with beforehand. The aim would be to determine whether an educational intervention is efficacious or not. The type of instruction we are interested in here is that aimed at the development of critical thinking.

The most immediate requirement for evaluating critical thinking skills lies in knowing whether a program -a program designed to teach people how to think- works or not. If instruction is given to improve certain skills, it is indispensable to know whether such instruction has any effect. To determine this, we must compare people's performance after the instruction with that observed before the intervention. If performance is better after it than

before it, we are on the right path to demonstrate that the later gain in skills is probably due to our instruction. This would be the origin of most projects aimed at evaluating critical thinking: to demonstrate the efficacy of an intervention. However this is not, or should not be, the only requirement of a measurement of thinking.

Critical thinking is attracting especial interest in higher education in countries such as the United States. There, some time ago a panel of experts studied "the future of higher education" and that commission suggested the generalized use, within the federal context, of tests on critical thinking in university students (Ennis, 2008). The intention of this initiative was set forth in the report entitled *A test of leadership*. The wish was to evaluate the competency of future leaders of society (U.S. Department of Education, 2006). The problem arising, however, is how to achieve this. It would seem reasonable to surmise that citizens who are going to have responsibilities in society in the future should have adequately developed abilities to argue matters out or make good decisions; that is, they should show good performance in the skills defining critical thinking. Social concern about leaders being able to make good decisions or solve problems seems perfectly reasonable, and we thus see that the insistence on measuring the capacity for reflexive thinking is not exclusive to investigators devoted to theoretical studies but is also a serious social issue.

Since what happens in the United States has greater influence than what might be suspected, that social concern is extending to many countries within the sphere of US influence, in the sense that initiatives are being made to assess the intellectual competencies of university student. One of the most generalized proposals is to evaluate the main skills of thinking; i.e., what is understood as critical thinking, such as the ability to argue properly, propose hypotheses, emit judgements about probability, and decide about or solve complex problems correctly. The interest shown by those responsible for higher education in several countries lies in their desire to ensure that apart from offering good training for a future profession university teaching should guarantee that suitable attention will be paid to critical thinking. From this perspective, tests on critical thinking would constitute a diagnostic and prognostic tool. In the former case (diagnosis), they would allow us to know whether an educational system achieves what is expected of it: the training of good professionals. If this objective is not met, it is possible to take measures so that the system will work better in the future. Thus, assessment serves to improve our educational systems.

Regarding predictive function, such tests enable us to know who has certain abilities that are necessary for assuming certain social responsibilities. Good judgement and the ability to make reasoned decisions seem to be the most appropriate demands of those who must solve problems that probably affect large numbers of citizens.

It therefore seems necessary to foster good thinking in people in general and in our students in particular. This makes it crucial to evaluate those good aptitudes in order to know whether they are present or not and to see which educational program develops them adequately. In sum, the aim is to determine which educational system, institutional or not, works; which intervention program fosters the skills that form part of what is known as critical reflection. This seems to be the departure point for the assessment of critical thinking; the need to know to what extent a group of people has acquired those skills.

THE METHOD OF EVALUATING CRITICAL THINKING

Despite the above, in order to know to what extent a person thinks critically there are at least two requirements: a clear concept of what critical thinking is and an instrument able to measure it. The reality of the situation, however, is very different, because we lack both a precise idea of what critical thinking is and a valid tool for assessing it. The question that arises from this situation is therefore whether it is feasible to perform an evaluation of this type in such circumstances. We shall leave this issue for the next section and focus now on addressing the problem in hand: the lack of conceptual clarity. And we say "problem" because there can be no instrument that is able to measure something that cannot be defined with precision, since we do not know exactly what it consists of. A test is developed to measure something when we know what it is, and it is there that its most important property lies: its validity; knowing what it measures. So before going any further let us attempt clarify what it is we wish to measure. Our efforts will focus more on stating what we understand by critical thinking and less on discussing the different ideas advanced in this regard (a good discussion of this can be found in Johnson (Johnson, 2008)).

As is always the case, the degree of conceptual precision is a reflection of our ignorance. When, for example, it is stated that "critical thinking is the intellectual activity that allows us to attain our ends in the most efficacious way" (Saiz and Nieto, 2002; page 16), we are not detailing intellectual activity, the actual processes. In fact, we are indicating that we do

not know what thinking really consists of. The above definition serves as a model, as an idea widely used in the field of critical thinking. However, let us try to obtain a more precise concept and let us say that critical thinking *is a process of seeking knowledge, through reasoning skills, the solution to problems, and decision making, affording us –with greater efficacy- the desired results*. In this second definition, we have specified the intellectual activity with an intrinsic goal inherent to all mental processes: searching for knowledge. Achieving our ends does not depend only on intellectual dimensions since we may require our motor or perceptive activity, such that it makes little sense to say that critical thinking allows us to achieve our goals since we can also do so, for example, by merely moving our legs. We must thus make an effort to identify the mental processes responsible for thinking.

More than with other things, thinking has to do with deriving something from something; inferring, reasoning. The main process consists of extracting new information, of seeking how to know more. We understand that what is essential about this process is inference or reasoning. Thus, reasoning is the core of thinking, but not only reasoning; we must refer here to the other two fundamental aspects of critical thinking to achieve our goals. We seek something when we don't have it, or when we do we want more of it or something better. This lack poses a "problem solving" situation.

Usually, we think to solve our difficulties. This is the second important activity of thinking. A problem can be solved through reasoning but also by planning courses of action or choosing the most appropriate strategy for the circumstances in question. Accordingly, to solve problems, as well as reasoning we must make decisions. Choosing is one of the most frequent and important activities we engage in when solving a problem. Therefore, we prefer to emphasize this aspect in a definition of thinking. Solving problems demands considerable intellectual activity, such as reasoning, deciding, planning… This latter characteristic (planning) goes beyond the actual mechanisms of inference. What we see when delimiting what thinking efficiently involves is that reference is made to concepts of different nature that go beyond what is fundamental to critical thinking: i.e., everything related to inference or reasoning. For this reason it seems opportune to speak of the components of critical thinking. Let us observe figure 1, inspired by the description provided by Halpern (2003) of a model of critical thinking

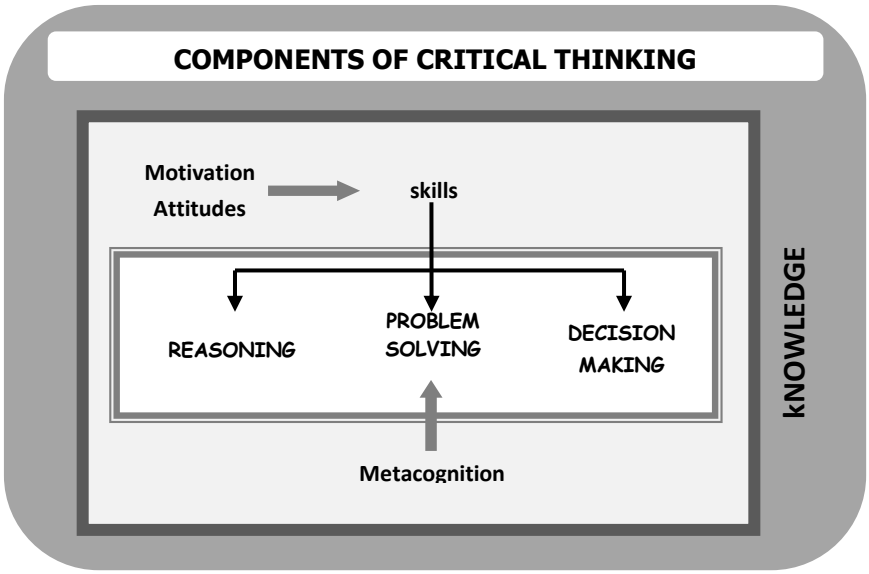

Figure 1. Components of critical thinking.

In this figure we can find all the concepts of the above definition but two: motivation and metacognition (attitudes are usually understood as dispositions, inclinations…., something close to motives, but at the same time to metacognition). The essential core of critical thinking continues to be reasoning, problem solving and decision-making, but why should we incorporate other concepts, such as motivation in a description of critical thinking? Some years ago it was confirmed that when speaking of critical thinking the fact of focusing only on skills does not allow all the complexity of such thinking to be addressed. The aim of the scheme shown in figure 1 is to provide conceptual clarity to the adjective "critical" in the expression "critical thinking".

Some authors consider this term unnecessary because they feel it to be redundant, since the critical character is consubstantial with the essence of thinking itself; it forms an integral part of it. Without allowing ourselves to be distracted by such nuances, we suggest that if one understands *critical* as referring to *efficient* one must also realise that this is not achieved only with skills. Other concepts must be joined to skills and must do so at different moments. Intellectual abilities alone do not achieve the efficiency assumed to underlie the term *critical*. First, such abilities to be set in motion, we must first wish this to be so ("knowing begins by wanting"). Thus, motivation enters the game before skills; it sets them in motion. In turn, metacognition allows us a direction, organization and planning for our skills in a profitable way and acts once skills have begun to be activated. Accordingly, motivation activates skills and

metacognition makes them more efficient. The goal should be a desirable knowledge of reality; more wisdom.

We believe that the fact of referring to the components of critical thinking, differentiating skills from motivation and metacognition, helps in the conceptual clarification we are seeking. On one hand, we specify which skills we are talking about and, on the other, which other components (other than *thinking*) are related to or even overlap them. We must be aware of how vain the illusion of finding "pure" mental processes is. Planning a course of action, an essential feature of metacognition, demands reflection, prognostication, choice, comparison and assessment… Is this not thinking? The different levels or dimensions of our mental activity must be related, or integrated. We believe that in this way integration will be better. Accordingly, our attempt to seek a conceptual clarification is directed towards that integration of the components of thinking. Our goal is to be able to identify what is substantial in thinking in order to determine what it is we can assess.

There is little doubt that thinking properly is profitable for our personal and professional lives. When we wish to achieve something or change a circumstance, the whole of our mind is (or should be) set in motion. We perceive the situation, we recall relevant aspects for the problem, we analyze all the information available together with that already known, we assess everything analyzed; that is we emit judgements about the most relevant matters, we decide on the options or solutions, we execute the plan, obtain results, which we evaluate; we estimate whether we have achieved our aim and, depending on the later degree of satisfaction after that estimation, we judge our course of action as being successful or not. The questions we must ask ourselves about this description is "where is the *thinking* process?": in all of that or only in part of it? Our answer is simple: we encounter *thinking* once we have all the information we need on the board game of life; when we begin to analyze it, from that moment we can say that we "continue" to think. More explicitly, inferring after the analysis, deciding, planning, using strategies…, all that would be *thinking*. As mentioned above, reasoning, deciding and solving problems are processes that we consider to be the fundamental skills of critical thinking.

We have thus defined what we understand by critical thinking and we have specified what it is for us (reasoning, deciding, solving). Our conception may be debatable, but at least it is not imprecise; it may be more or less correct, but it is not vague. Our wish to pinpoint what exactly critical thinking is is imposed by our *way of assessing it*. We cannot assess when there

is imprecision. *Knowing what one is measuring* is what legitimises a test designed for such a purpose. Everybody knows that assessing any mental activity is a difficult and frustrating task and in the case of a type behaviour with such a multicomponent nature as thinking even more so. Once we know what we have to assess, things seem to get simpler, although not that much. One only has to look at some sources to see this. It is not our aim here to carry out a review of the issue, although we shall evidently refer to the most important ones. In a book edited by Daniel Fasko, there are several chapters addressing the assessment of critical thinking that show how very difficult it is to measure what is mental (see Ennis, 2003; Halpern, 2003b).

From the pioneer work of Donders (1868/1969) to the most recent works (see for example Possin, 2008), investigators have always felt the temptation to quantify even the most slippery of entities, such as mental activity. To accomplish this, as we have said previously, investigations have focused on *how* and *what*: what to measure and how to do so. We have already specified the *what* and now we must tackle the *how*. Perhaps the author who is most knowledgeable about critical thinking is Robert. H. Ennis. Ennis has published the greatest number of tests and has conducted the best studies of the topic (Ennis, 2003, 2008). In the work from 2003, for example, he offers no less than 7 reasons for measuring critical thinking, and presents and assesses all the tests published. Of the 21 cited by that author, 4 (20%) were compiled by himself and his colleagues. With such a large offer of ways to measure it one could imagine that the way critical thinking is measured today would be correct, but nothing is farther from the truth. Most tests fail because they fall into the trap of objectivity. Nearly all of them consist of questions (excessively formal) to which subjects must respond by choosing one or the three or four options proposed. However, these tests have one virtue that is difficult to resist: the speed with which they can be corrected. But a test with a multiple-choice format and an artificial content will hardly measure the main skills of critical thinking. A set of items against which one must merely signal the one that seems correct demands little thinking and a lot of perception: it makes one discriminate, but not justify; it makes one choose, but not assess. In sum, it prevents what is consubstantial to the nature of thinking from being expressed. This manner of proceeding has as a consequence the three essential limitations of the tests published: those affecting *validity* and *production*, and *their complexity*.

In order to measure the reasoning capacity – to mention the most significant aspect of critical thinking- of a person taking a test, that person must make his/her mental process explicit so that, from his or her manifestations, it will be possible to infer how s/he is reasoning.

If when challenged with a problem a person simply has to decide if the solution offered is correct or not, or choose one of the replies offered, it will never be possible to gauge what type of reflection has been engaged in for her/him to respond. One can only speculate about this. In the case of a syllogism such as "All men are mortal. John is a man and therefore John is mortal", if we only have to say whether it is correct or not there is no way of deducing how that answer was reached. Several studies have shown that by guiding ourselves by quantifiers it is possible to develop strategies that guarantee up to 80% efficiency without any knowledge at all of logic. The *atmosphere effect* is a trick that allows such efficiency to be achieved (Anderson, 1990). The person responding to the problems follows simple practical rules, such as seeing which quantifier is repeated in the premises and in the conclusion. In this way, it is possible to attain a high number of correct answers, hitting the target without thinking about the problem. The impossibility of knowing the thinking process followed by the person answering is perhaps the most serious problem in most existing standardised tests. And this is so because it detracts from the test's validity.

The second severe limitation of the tests published has to do with the type of answer sought. If we use a multiple-choice format, as well as the above limitation to validity another one emerges: we are only asking for an *understanding* of the task, but not the *production* of thinking. Assessing thinking consists of evaluating what is produced, not what is *assumed* to have been produced. With a test involving alternatives, we assume that thinking is generated, but we have no proof of this. We cannot quantify the process because it has not been made explicit. Paradoxically, one is trying to measure something that has not been expressed. In a test of this nature, we should be asking subjects to express *why* a particular way of answering is used; we should be asking for an explanation of all the replies requested and given. The problem, it may be said, is the short test correction time required. What does a test that is corrected mechanically serve for if we do not know whether it is measuring what we want it to?

The last limitation we shall address refers to *complexity*. The previous example "All men are mortal…" illustrates this problem quite well. A trivial task, excessively formal and artificial, de-motivates and fails to activate the usual mechanisms of thinking that we use in our everyday lives. And if this happens, then we are evaluating answers to a test that reflects little skill, because what could be happening is that the test is not adequately activating the pertinent thinking process. Most tests published also suffer from this third limitation. To check to what extent reflection about a task is being engaged in, it is important to propose problem situations

similar to those encountered in daily life; it is recommendable to "simulate" quotidian difficulties in the test; we must propose problem situations like those found in daily life. The problems of standardised tests do not pose daily situations to be solved. Again, we see an important deficiency that prevents us from assessing what we are really trying to measure. If a person doing the test is not at all interested in the exercises, that person will answer without becoming involved intellectually, without setting in motion the reflective machinery that we are trying to capture. If the person does not think with determination we shall never be able to evaluate such thinking.

The way in which most tests evaluate critical thinking is inadequate. With a few exceptions -to be discussed below- tests with a multiple-choice format cannot explain conceptually what it is we are trying to assess, because the type of answer demanded prevents this. We can play, and very much so, with statistics and factor analysis and observe clusters that are more or less close to the starting premises, but we will never be able to say that those "factors", those clusters, are produced because the appropriate thinking process has been activated. We cannot affirm this simply because the type of answer is unable to demonstrate it.

Asking someone who is taking a standardized test to "mark" what seems right prevents that person from really thinking about the issue. If the test format limits what we wish to measure, it seems reasonable to assert that something is being done wrongly. The idea is not to measure thinking and at the same time prevent, or limit, people from thinking. Using stereotyped or "overly formal" problems in the tests invites the person taking the test to become distanced from it. This distancing from the proposed problems certainly does not stimulate deep reflection, which is what we are supposed to be after. We said earlier that there are some tests that are an exception to all this, because they do not have the limitations just discussed. The problem of these instruments, however, is that they only measure one aspect of critical thinking. For example, the *Ennis-Weir Critical Thinking Essay Test* (Ennis & Weir, 19895) assesses well, but it only addresses the capacity for argumentation. There are other single-dimension tests that can be consulted in Ennis (2003). However, what we need is to measure all the skills of critical thinking, not just one.

VIABILITY OF THE EVALUATION OF CRITICAL THINKING

Most standardized tests that we have evaluated, with their important deficiencies, paint a rather dismal picture of the possibility of assessing critical thinking. The viability of a project designed to measure this type of thinking seems to run into problems that are so serious that they may dissuade even the most optimistic investigator from proposing a task of this kind. It is true that the evaluation of complex processes such as those of critical thinking poses discouraging challenges, even for research teams equipped with excellent human and material resources. However, the greatest difficulties do not usually lie in a lack of resources but in a lack of ideas. When someone opens the door to the solution with a suitable strategy, many difficulties that were previously believed insurmountable disappear immediately. One measurement project, developed by Halpern (2006), has managed to achieve precisely this: to open the door to many of the dead-end streets plaguing the evaluation of critical thinking. In our opinion, the HCTAES test (*Halpern Critical Thinking Assessment using Everyday Situation*) marks a *before* and *after* in the evaluation of critical thinking. Its contribution lies not so much in the quality of the tests as in the proposed assessment model. The test has been adapted to the Spanish by us and studied in sufficient depth to be able to state that it should be improved in certain crucial areas (Nieto, Saiz and Orgaz, 2009, in press). It is precisely these deficiencies of the test that prompted us to develop alternative instruments, described below. Before this, however, what we are pursuing is to study whether it is possible to evaluate critical thinking with a certain degree of efficiency. And what we must do before anything else is to demonstrate the *viability* of a project aimed at evaluating critical thinking.

As explained above, there are at least three severe problems in most of the standardized tests published: the first is that we do not really know what they are measuring; the second is that they do not activate the main skills being studied and, finally, the problems posed are artificial and divorced from everyday reality. In contrast, the HCTAES test attempts to circumvent these deficiencies, although the results do not seem to support the notion that this has been fully achieved. The contribution of the test is not so much its final product, which can be improved, but the novelty of offering a decisive proposal that does make it viable to assess critical thinking. This is its very great merit. And it is this circumstance that has guided us in our own assessment project. We shall begin by describing how the problems of validity, production and complexity are resolved in this test. Then, we shall analyse what it is able to measure, and finally we shall explain how our project tries to solve the aspects in which the test

fails. In our assessment, we take up the HCTAES measurement model, but correcting its limitations and incorporating a new proposal that we believe solves the most important problem of the three: i.e., *validity*. To resolve this crucial limitation of any measurement, we pose a design of the items based on *task analysis*. However, to understand our proposal other territories must first be explored.

Let us take an item from the HCTAES test to help in the analyses performed below.

Item 21, Part 1 of the HCTAES test (Question 41)

> Suppose that you are a first-year student at a dental school. You realize that your new friend, who is also a first-year student in dental school, is getting drunk on a regular basis several times a week. You do not see any signs of her drinking problem at school, but you are concerned because you will both begin seeing patients at the school's dental clinic within a month. She has not responded to your hints about her drinking problem. As far as you know, no one else knows about her excessive drinking.
> (A) State the problem in two ways.
> (B) For each statement of the problem, provide two different possible solutions.

This item of the test is a good illustration of how well the test works, as we shall see. The test has two parts: one open (as in the example) and the other closed, in which the same situation is presented with several options, from which the subject must choose one. The importance of the approach of the test is found in the open part, and so we shall now address this (for information about the instrument, see Halpern, 2006, and Nieto, Saiz and Orgaz, 2009, in press). This tool is original in nearly all its aspects, both as regards the nature of the items and in the way of correcting them. In the 25 situations included in the test, problems involving the main thinking skills are posed. In particular, the following situations are presented: 1) checking hypotheses, 2) verbal reasoning, 3) argumentation, 4) probability and uncertainty, and 5) decision making and problem solving. Situation 21 would be of the last type –decision making and problem solving. The items included in the test are daily problems that are replied to, at least once, in an open fashion and they pose problems that must be solved via those five main skills of critical thinking. If a problem is presented that must be solved with a specific mechanism of thinking, robust validity of the test is ensured. If, additionally, the subjects are asked to explain their answers to each item, we are obliging them to produce thinking. Finally, if we pose problems that resemble everyday situations, we are achieving a primordial proximity

to and including the *complexity* involved in the tasks and problems addressed in daily life. Item 21 is an example. In that item, we must answer using a problem-solving strategy and we must also justify the solution we propose. We are solving a daily problem. This test is very ingenious because it addresses, in a simple way, the three main problems all at the same time. Let us study them one by one.

A situation that poses a daily problem means that the person doing the test is familiar with what is being proposed, and hence that person will tackle the problem as usual, thinking or reflecting as s/he normally does in everyday life. This has at least two clear advantages: one is that people will become involved in the task and the other is that they will use the best skills available to them. Such tasks motivate more than others because they are similar to what we encounter day by day as regards *complexity* and importance. No artificial problems outside the usual sphere of interest of people in general are posed, and neither are any problems that might distract them by leading them to seek ways of solving the problem that they do not usually use. With these items, we at least manage to achieve a better motivation and intellectual activity in the people solving them. People become deeply involved in the task. Concerning this, both we and the author of the test have data concerning its application that validate this involvement. We believe that it is important to use tasks that are familiar to people; tasks that allow people to become involved easily and interestedly. The ecological validity of a test is important, but so is the *generalization* or *transfer* of skills. This concept has not been mentioned up to now because it lies within the remit of both intervention and assessment. And here we are especially interested in the latter. If in the assessment we present a problem that is very different from what is encountered on a day-to-day basis, we run the risk of weakening the validity of the test. If with an item we are trying to get people to use a given skill, we must facilitate as much as possible a situation in which that skill, and no other, will be used However, if the item poses a problem quite different from the kind we usually run across, two possibilities arise when we attempt to solve it: one is that we shall solve it without using the skill required, and the other is that, having the skill, we don't know how to apply it owing to the differences between daily situations and the situations of the task in question. This latter possibility, if it arises, prevents a test from assessing what it is supposed to, simply because it fails to activate the thinking mechanisms it is supposed to capture and for which it was designed. It should be understood that the ecological validity strengthens the theoretical validity, also known as construct validity. It is not merely a question of the task being more interesting to the person

performing it, (which is also the case) but that it must manage to set in motion the mental processes it is trying to measure. It would be impossible to measure something that that is not there to be measured. This aspect of transfer is much more important than what can be explained here, but it is extensively addressed in Saiz and Rivas (2008). The first great virtue of Halpern's work is that she rightly poses problems that are close to people's experience. If we analyze situation 21, we see that it is a kind of "simulation" of problems common to our daily lives. An event is described that we must place in situation and for which we must propose solutions. Since we cannot create real problems, we simulate them and ask for solutions. Let us make a comparison. Here we proceed in a way similar to what happens with future pilots: before they are allowed to fly they must spend many hours in a simulator. A simulator is not a plane, but is certainly as close to one as possible. We do not measure how real problems are solved (evidently, we fear that they would not allow us into their lives); only simulated ones. Daily situations are problem tasks that stimulate all the above described aspects and this simulation is so important that it is picked up in our proposal, described below.

In situation 21 the following are requested:

(A) State the problem in two ways.
(B) For each statement of the problem, provide two different possible solutions.

In this test, an open-response format is used, evidently linked to the content of the items, as seen in situation 21. The person doing the test must explain what he or she is doing. In item 21, we are asked to define the problem and offer solutions. We are asked to think about the situation; we are asked to produce thinking. Problem situations make us reflect and -what is of huge importance for assessment- express that reflection. Halpern's test opts for two response formats: one open, in part 1, and the other closed, in part 2. Regarding the open format, this is necessary for the assessment of critical thinking, since it is not possible to measure such complex skills in any other way. A closed-response format (multiple choice, for example) prevents us from expressing our thinking, which prevents measurements from being made. There is little to say about this aspect; we can only measure what can be seen. The HCTAES test is particularly sensitive to this problem, despite the high time cost involved in the correction of the open responses. By contrast, standardized tests sacrifice validity in an attempt to save time. This is simply nonsense. Open-response formats offer an additional advantage, no less important: when a task involves producing something, as well as what has been mentioned

above it demands a more complete manner of thinking than when only understanding is demanded. A closed-response test merely measures understanding. In a test demanding production, however, we must elaborate on what we have thought for it to be intelligible before we express it. We must re-think what has been thought to express what we are thinking. Everybody knows that having an idea is not the same as communicating it. When we communicate we often find that the idea was not as clear as it might have been; when we communicate we see shadows that were not present in our minds during our analysis. This feature of the task (production *vs.* understanding) is important in other contexts because it means that depending on the nature of the context, certain biases escape ourselves from our main analysis, the atmosphere effect mentioned above only appears in tasks of understanding and never in those involving production (Shaw & Johnson-Laird, 1998) The aim of this discussion is to provide further support, if that is possible, to the importance of allowing what is being thought to be expressed in an assessment task. This is essential from any point of view. Again, Halpern's test fulfils this requirement very well.

Two of the three problems mentioned in the previous analysis are well resolved by Halpern's test. The core problem, however -that of validity- is not. Let us see why. The test claims to measure the 5 previously mentioned factors (hypothesis checking, verbal reasoning, argumentation, probability and uncertainty, decision making and problem solving). Item 21 is an example of the last factor. The situation picks up on a daily problem that matches the factor proposed quite well, but how is that item corrected? In the response protocol the following appears:

RESPONSE PROTOCOL

(A) State the problem in two ways. (B) For each statement of the problem, provide two different possible solutions.

(0 to 6 points possible)

1 point for each statement of the problem. (2 points possible)
2 points for each solution provided. (4 points possible)

In order to get full credit (6 points):
The problem must be stated in two distinct ways.
There must be two distinct solutions for each problem identified.

(A) State the problem in two ways.

She has a drinking problem and will be dealing with patients.

She has a drinking problem but doesn't show signs that it impairs her ability.

(B) Provide two possible solutions for each statement of the problem.

She takes steps to curb excessive drinking.
She does not deal with patients because of the problem.

2) You show her how it is impairing her ability or show her how it could.
You convince her about how she is putting people in danger regardless of if she knows it or not.

Examples of common problems:
The friend has a drinking problem.
Patients could be harmed by the problem.

Examples of common solutions:
Talk to the friend about the problem, find a way to prevent the friend from drinking.
Keep the friend out of the clinic; talking to the friend.

If we read the response protocol carefully, we see that redundant expressions of the problem are being scored, as in the examples. Giving a point for saying that she has a drinking problem and that this could affect her patients seems to be too generous, above all when we are trying to assess thinking, not perception. The same could be said about the solutions. Suggesting that she should drink less and award a point for that is also too generous. What are we assessing? Defining a problem properly and solving it efficiently? Nothing that is proposed as being worthy of points fits in with this. This problem is encountered in nearly all the 25 situations of the test. The test offers tasks that can tap the main skills of thinking well, but it fails when attempting to assess ways of answering that, in many cases, have nothing to do with such skills. It contemplates answers that are not in the context of what one is trying to measure

and no single correct answers are proposed. Since there are no exclusive ways of solution, the validity of the test is diluted.

Govier (1987) highlighted the limitation of thinking tests when she indicated that the correct answers and a high percentage of the answers of the person doing the test must be the same for us to be able to say that that person has the thinking capacity required to answer correctly. If we fail to design a problem that demands a single reply our attempt to measure a given skill will fail miserably. The most important limitation of the reference test lies in the wrong way of correcting it, although not only in this, since many of the situations posed demand a more careful specification of the problems. In our opinion, the main limitation of this test is the fact that it has not designed daily situations in such a way that they can be answered with only one correct reply. This has serious consequences as regards validity that we shall discuss below, at the same time offering a solution.

Let us analyse this issue in a simple way. Let us consider that we wish to measure casual reasoning. To do so we build a problem of this nature that can be answered in only one way possible, because in that way, when the answer is correct, we can infer that a specific way of thinking has been activated. This way of proceeding would be the one needed for solving problems of causality and it would not be possible to solve them any other way. We require situations whose strategies for obtaining a solution are mutually exclusive, such that the one we use to solve the first situation cannot be the same as for the second one. In this way we shall be able to be sure that the causality item is only measuring causality. Only if we can achieve this specificity of problem solutions can we measure what we are seeking to measure and thus make our test valid. Halpern's test fails in the terms stated because it does not perform a proper analysis of the tasks it proposes. Let us see this in greater detail.

As mentioned above, there has always been interest in measuring mental processes. Some time ago, Donders stated something that was very interesting: "Shouldn't thinking have a finite duration and might it not be possible to determine the time required to form a concept or express one's own wishes?" (Donders, 168/1969, p.147) This author inaugurated what would later be known as *mental chronometry*. For our aims of assessing thinking, what is especially relevant is not so much the measurement of the duration of mental processes as an analysis of the procedure that must previously be implemented to be able to measure them. Task analysis

methodology is born of this effort to anticipate which "mental steps" must be put in play to successfully perform a given operation.

In order to understand this technique better, let us summarise the reflections of Donders; let us comment on the task that author created. In the test, the experimenter uttered syllables and the subject replied with the same syllables. Task (a) consisted of the investigator pronouncing only one syllable, which the subject had to repeat. In task (b) one syllable out of six possible ones was pronounced and the other person repeated it. In task (c) a syllable was fixed that was the only one that had to be replied to while the others were not to be repeated. The correct performance of each of the three tasks demands at least three processes: 1) Stimulus coding; 2) response decision, and 3) response emission. However, for task (b) it is also necessary to discriminate among the stimuli and decide which response to emit. In contrast, for task (c) it is only necessary to discriminate among the stimuli but not to decide what to reply. What Donders proposes is really ingenious: that author uses three tasks to activate specific processes. In this way, by comparison it is possible to assess specific processes of discrimination or motor decision.

What is interesting about this technique of analysis are its potential to solve the problem of validity. Halpern's test only goes half way to solving the problem because it does not apply this method. As we saw previously, daily situations can elicit a single thinking process but, in practice, the test fails to do this because in its design the reply that is supposed to be given is missing. In all Donders' tasks, the stimuli and the replies that must be given are fixed. In contrast, in Halpern's test only the former are determined; that is the problem situations. The correct answers are not fixed for each of these situations, such that we cannot know what the appropriate strategy for replying correctly is and neither can we know how to correct the answers consistently.

Our proposal is to design a test with daily situations that define specific problems for which there is only one correct answer. In this way, upon detecting correct answers for each problem task we shall know unequivocally which mental process is being used so that we can quantify it more reliably. With this specificity in the triggering of concrete thinking mechanisms, the serious problem of validity is overcome, such that we will have an instrument able to measure the thinking skills we wish to measure and no more.

Returning to the example of situation 21, to improve this task it would be necessary to design the problem in a clearer way with the aim of ensuring the existence of only one correct way of answering it. It would be necessary to propose a task that can only be addressed in one way and that can also only be solved in only one way. This would allow us to quantify the replies as though they were dichotomic variables: correct/incorrect. Additionally, another no less important problem is solved; that of the lack of reliability of open answers, because in these the score depends on the subjective criterion of the corrector. In our case, this problem does not arise because there is only one way of scoring the answers however open they are: the ones considered correct are those that solve the problem and those that do not are rejected as incorrect. However, this will be seen clearly when we explain our proposal.

PROPOSAL FOR THE EVALUATION OF THE MAIN SKILLS OF CRITICAL THINKING

Our enquiry into the assessment of critical thinking has allowed us to identify certain major problems of measurement and to obtain a glimpse of a way to solve them. The desire to use closed-response formats and artificial tasks in most standard tests published renders them unable to measure what it is they were designed for. If a person is asked to solve reasoning problems simply by marking the options given, that person cannot generate or produce his/her own replies; the person is obliged to choose from replies that are already there. This type of test neither stimulates subjects nor allows them to express what they are thinking; this prevents us from assessing what they are thinking. Also, if the tasks we use for such an assessment are very different from the ones people encounter on a daily basis they will not result in an incentive to reflect but rather in an invitation to passivity and lack of involvement. In sum, if we do not offer problem situations that activate and exteriorize the processes that we wish to measure, any project aimed at assessing critical thinking will be a failure as from its very inception.

The HCTAES project is a proposal that attempts to solve these difficulties. The results do not seem to have fulfilled this aim in a completely satisfactory way, but its approaches did inspire us in our own work. The test solves the problem of artificiality and production quite well. It offers daily tasks that induce the person doing them to become involved and allows them to confront them, activating the operations that must be followed to perform them. The person doing the task has the freedom to comment on problems similar to those that s/he would

encounter in every day life. However, the test does not close the way of answering in only one sense. The problem situations proposed do not offer a unique way of answering because they do not elicit the use of just one process. As we have seen, if each task stimulates more than one thinking process and more than one way of answering, we lose the chance to see which particular process has been engaged in to achieve the correct answer. Under such circumstances, it is no longer possible to know what we are measuring, simply because when challenged by a given task it is possible to think in different ways. There is no longer the possibility of measuring what we are seeking to measure, and hence no validity.

One way of fixing this drawback is to design problems with only one right answer; this indeed is our proposal. We take advantage of all the virtues of Halpern's proposal and correct its deficiencies. To do so, we incorporate task analysis into her proposal. The daily situations of the type presented in the HCTAES test can be redesigned in such a way that they pose problems that only admit one answer. Following on from the wise counsels of the founders of "mental measurement", we design tasks whose resolution we can imagine beforehand. If to solve them it is necessary to follow step A, then step B and finally step C, for example, when a person solves the problem properly we can be sure that that person did in fact follow those steps in that order and not in any other. Thus, we can at least know that when a test problem is answered correctly it is because those mental processes (A, B and C) have been activated. When we measure performance or efficiency we will know that we are addressing only those mechanisms. In this way, the huge problem of validity is solved.

Upon analysing the characteristics of a task, we can imagine the operations that will be required to solve it and this ensures that we will know what we are measuring; additionally, this will allow us to work in the opposite direction. With this method, we are not only proposing a problem for which we can imagine the steps that would be taken to solve it, but also we can imagine which steps we wish to be taken; then we can design a situation for which those steps must be taken. The advantage of this reversible nature of task analysis is seen when elaborating measurements of thinking, since this characteristic will allow us to make the problem tasks compatible with what we wish to measure. In our case, the main skills of critical thinking, described at the beginning of this contribution in turn consist of other more specific ones. These more concrete skills of critical thinking are what we are really assessing and the tasks are designed specifically to measure them. One of the major skills is reasoning, but there are many forms of this such that we must select, with sound criteria, those that best represent the capacity

and measure them (Saiz and Rivas, 2008). By following this route, what we are doing is fixing the structure of the assessment test, to be described below. It is now necessary to establish how we are going to operativise that structure, because this is another part of our work: choosing relevant thinking processes and creating specific tasks for each of them. Our assessment proposal guarantees the validity of the measurements through this manner of analysis

The need to measure thinking was justified at the beginning of this contribution. We stated that one of the reasons for measuring this type of process is to estimate whether an intervention will be efficient or not. For some time now, we have been developing and trying to improve an intervention program (Nieta and Saiz, 2008; Saiz and Rivas, 2008). A program is efficient when: a) it produces a change, b) this change persists with time, and c) it is generalized or transferred (Saiz 2002). The permanence of a change demands that measurements be made after some time has elapsed. Efficiency demands the use of a test that fulfils the conditions established above in our analysis. And this is what we shall develop: a test that will correct the deficiencies that others have been unable to resolve. Our test is called PENCRISAL (from: ***Pen****samiento* (thinking), ***Crí****tico,* ***Sal****amanca*). Finally, transfer is the most important measure of efficiency because it evaluates people's ability to generalise skills to different domains, once those skills have been acquired or developed, because we cannot transfer something that does not exist. This is why it is the most important measure: it assumes a prior change in our thinking processes. This is the second test, and it is called PENTRASAL (***Pen****samiento*, ***Tra****nsfer*, ***Sal****amanca*).

Before describing our two tests, it is important to note what is meant here by *transfer*. Our assessment proposal consists of using daily situations that will guarantee the theoretical and ecological validity of the test. When people engage in a test aimed at assessing thinking, they are required to use the skills deployed in daily life. For this reason, tasks similar to daily situations are designed. This approach is the one we maintain in the test for evaluating the effect, such that now the issue is: What is the difference between the two tests? Why does one of them measure the magnitude of the effect and the other one transfer, when the design of the items is identical in both cases? The answer lies in one of the goals of our intervention; namely, to measure the efficiency of an intervention program. Our approach to the intervention determines our transfer measurement (for a more detailed discussion of this, see Saiz and Rivas, 2008). Here it suffices to clarify the following. An intervention supported by tasks similar to those encountered in daily life cannot seek transfer in those contexts because those

tasks are the ones that have been worked on previously. We must seek transfer to daily domains; i.e., contexts that cover all spheres of our daily activities. This is the main difference between the two tests. In the test assessing the magnitude of the effect, the tasks deal with personal problems, while in the transfer test the tasks deal with problems from different contexts, such as education, health, job-related problems… All this will be detailed below. The idea, then, is to see whether it is possible to solve problems with the same efficiency in all the spheres of our personal life. This would be our measurement of transfer and it can be consulted in Saiz and Rivas (2008). Let us now discuss our assessment proposal.

As stated above, the development of our tests arose, on one hand, from the need to check the efficiency of an instruction program that has been under development for some time (Nieto and Saiz, 2008; Saiz and Rivas, 2008) and, on the other, from the need to have truly valid tests that will measure critical thinking. Accordingly, to overcome the problems described above and at the same time check the validity of our instruction program in its dual aspects –magnitude of the effect and transfer- we have developed two specific measures: PENCRISAL and PENTRASAL.

With these instruments we can assess whether the instruction has afforded the expected changes; that is, whether there has been an improvement in critical thinking, and also whether a generalized mode is being used: i.e., whether those skills have been transferred to other domains.

To construct the tests we followed the conclusions derived from our previous analyses, these leading us to adopt four main principles: 1) pose items that are problems found in daily situations; 2) use an open-response format; 3) propose problems dealing with different topics of knowledge and domains, and 4) pose problems with only one answer.

One of the central aspects of our approach is that it refers to the use of problem situations that are as close as possible to those found in everyday life. The use of such situations, which reflect real problems, on one hand allows us to assess our skills in thinking in a context similar to what might be found in real life and, on the other, to introduce an important motivational factor, achieving greater interest in the task proposed.

Let us consider the following examples:

Example 1.

Item 16. California Critical Thinking Skills Test

> "If Alex loves someone, it is Barbara he loves. There are many people who Barbara does not love, and Alex is one of them. But everybody loves somebody"

Example 2:

Situation 1. PENCRISAL

> "John's personal trainer has told him that only if he trains for two hours a day will he pass the exam to enter the Fire Department. John is worried because he hasn't trained for the requisite two hours every day and so he believes he will fail the exam"

A deductive problem such as example 1 has too formal a structure, which distances it from our daily doings in which we might need to know this type in principle. However, using tasks such as example 2, we move away from the more formal approaches in thinking and come closer to more ecological contexts, which will allow us a broader application of the skills in daily situations.

In this context, the question that arises is: What topics do the problems deal with? It is at this point where we must differentiate between the two tests presented, which we have explained previously but which we now further specify as regards their content. The items of PENCRISAL, which as stated before is a measurement of the magnitude of the effect, mainly refer to a single context of knowledge: that of *personal relationships*. In contrast, those of PENTRASAL, the measurement of transfer, refer to all spheres we consider to be important in our daily activities, such as health, sports, leisure, traffic contexts, education, the environment, politics, work-related issues and consumption.

For clarity's sake, let us recall how we justified this distinction as regards the concept of transfer. If the difficulty involved in generalising our ability to reason or solve problems lies in the difference between the domain of learning and daily life, and if we pose an intervention with daily situations to reduce that distance, what can be seen from this is a reconceptualization of the idea of transfer. Let us recall that transfer is the ability to generalise a skill from one domain to another. We now find that the domain is the same; what we introduce in this test are the contexts of knowledge of problems, such that if we are able to apply our skills to different spheres of knowledge, then we can continue to speak of transfer.

Another of the relevant aspects in the construction of these instruments is the use of an open-response format, whose advantages over a closed format have already been discussed. Nevertheless, let us return to some of them. The latter format provides insufficient information about the thinking process(es) that the person must presumably set in motion to solve the problem. These are answers that demand the recognition of options, but not their production. In contrast, the items of our tests, which feature an open-response format, allow thinking to be produced and enable us to know the specific mechanisms or processes involved in that thinking, Items of this kind allow subjects to generate and explain their answers, such that they must not only identify those situations in which it is necessary to activate thinking processes but must also decide which of the processes is necessary for solving the particular situation in the most suitable way. Below are some examples.

CATEGORICAL REASONING ITEMS

Example 3. Closed-response format

Cornell Critical Thinking Test: Item 6

"MR. Wilstings has proposed that we open our doors to all the foreigners who want to enter our beloved country. But foreigners who want to enter our beloved country. But foreigners always have made trouble and they always will. Most of them can't even speak Enghish. Since any goup that makes trouble is bad, it follows that foreigners are a bad bunch".

Please answer as follows:

If the conclusion *necessarily follows* from the affirmations, mark A

If the conclusion *contradicts* the affirmations, mark B

If the conclusion *neither follows* from *not contradicts* the affirmations, mark C

FORMULATION		REPLY
All Foreigners (F) cause problems (P)	F→ P	Option A
All groups that cause problems (P) are bad (M)	P → B	
All foreigners (F) are bad (B)	F → B	

Example 4: Open-response format

PENCRISAL: Situation 5

The function of understanding involves language and only humans use this, such that it may be concluded that only humans understand.

Is the reasoning correct? Explain why.

FORMULATION		ANSWER
Understanding (U) involves language (L) Only humans (H) use language (L) Humans (H) understand (U)	U→ L L→ H U→H	Correct reasoning *Explanation*: Transitivity principle

In these two items, the situations are equally well posed. As may be seen in the tables, it is possible to suitably extract the formal structure of both arguments. However, what we wish to highlight is that, despite this, what is provided by a closed-response format, such as the case of the Cornell item, when it is only necessary for the subject to identify that the correct answer is A, is very obscure information, since the very fact that someone has been able to identify that option as valid does not offer information about the thought processes involved. It is merely putting into motion strategies of identification and recognition. In contrast, in the open-response format (example 4) on one hand the subject correctly identifies that the argument is correct, and then puts the production processes in motion because s/he must make the explanation explicit, and it is here where, in order to state that it must be due to a transitivity principle, the subject must use specific thinking processes.

The last of the four basic principles used in our tests is the design of problems with only one possible answer. Let us recall that this proposal is the one that allows the serious problem of validity affecting most tests to be solved, since when a person is challenged with a problem task concrete processes are activated that lead to concrete responses. We elaborate tasks that elicit specific thinking processes that will always lead to the same answer. In this scheme, when that unique answer appears we know which particular mechanism was set in motion. Thus, we can be sure that we are indeed measuring what we want to and, additionally, we simplify and bring rigour to the correction of the test. Perhaps our approach could be better understood if we compare two items: one that does not guarantee validity and the other that does. One of them is Halpern's and the other one is ours

Let us see some specific examples of the HCTAES and PENCRISAL.

PROBLEM SOLVING ITEMS

Example 5: HCTAES (Halpern): Item 24 (47)

Suppose you are caring for your neighbor's dog and one task that you have to do is to get the dog to take a large, apparently bitter, pill. The dog is a large attack dog that bit a small child last year. How would you go about getting the dog to take its medicine?

List two good solutions to this problem.

CORRECTION
2 points: if the subject uses creative methods 2 points for answers similar to*:"ask for help, consult an expert, pill with something", etc.*

Example 6: PENCRISAL: Situation 22

You are the owner of a family oriented bar where your mother works in the kitchen and your daughter sometimes works as a waitress, You now need to hire another hand to keep the bar open because you mother has been admitted to hospital and the doctors have told you she will be there for some time. Also, you daughter has begun studying at a University outside your town. You must select from among several candidates and you don't want to ask your daughter for help in this because she would have to make a trip. A friend tells you he has a daughter who is out of work and needs a job, but you have your doubts since she has never worked in the catering industry.

What strategies or steps should the owner follow to reach a good solution? Indicate the resulting solution.

CORRECTION
2 POINTS: IF A METHOD IS APPLIED WELL Two points for answers that include: - *Identification of the problem :* Neither the wife nor the daughter work in the bar and the owner needs another member of staff - *Subproblems or subgoals*: ✓ To make a selection of candidates to choose the best worker , which will take sometime and possibly make it necessary to close the bar ✓ Provisionally contract the daughter of the acquaintance, in the meanwhile searching for the best candidate.

In example 5, two points are assigned to solution that, according to that author's criterion, are creative methods of solution. However, in our opinion if we are really trying to measure a person's ability to solve problems, it does not seem that "*ask for help*" or "*consult an expert*" imply that some kind of procedure is being used, neither as regards solving problems nor with respect to thinking. We could instead consider these answers as "irrational" since they do not activate any kind of thought process at all. In contrast, in example 6 (from our test) it is possible to appreciate an initial difference with respect to the previous one in the question posed about the situation. On asking "What strategies or steps should the owner follow to reach a good solution? Indicate the resulting solution", we are making explicit both the demands of the task and the procedure through which that solution is reached.

Another evident difference is the very precise delimitation made of the types of answer to which we are going to assign, in this case, two points. As stated above, the problem situations were designed following the method of task analysis, through which the problems are posed in such a way that we can anticipate the operations that people will need to apply to solve the problem. And this is possible because, as said, there is only one way of solving it.

Let us now devote some time to a more in-depth description of the most relevant characteristics of our two tests. We shall address them jointly since they are based on the same fundaments and have a similar structure.

The construction of the tests started by compiling a database of items sufficiently broad to be able to make a good selection, attending to the level of difficulty of the items, that would be sufficiently heterogeneous to be able to make a precise assessment of the different skills. Additionally, this set of items guarantees, as far as possible, the content validity, reflecting the main aspects representing the skills of critical thinking.

In the choice of problem situations, two aspects were taken into account: first that the thought processes or operations to be assessed would be well represented, and second, that their written presentation would be clear and presented in a colloquial tone, avoiding technical language.

The first version of the tests was applied to a sample of 469 university students from different disciplines with a view to performing the corresponding psychometric analysis of the items. Based the results of this application, some items were removed and replaced by others, and those that showed any slight deficiency were also eliminated. All this aimed at improving

the quality of the instruments. From our results, it may be inferred that the tests show sufficient reliability and a validity in consonance with the theoretical postulates. As this paper is being written, we are beginning to conduct the empirical application of this second version, with the confidence that the improvements introduced will be demonstrated psychometrically.

Currently, PENCRISAL and PENTRASAL comprise 35 items (problem situations) with open answers. The contents of the problems were developed in such a way that they do not require the answer to be elaborated and expressed in technical terms but in colloquial language. This can be seen in the following example.

Example 7: PENCRISAL: Situation 20

> John needs to use public transport every day to get to work and the trip takes about two hours. In recent days, with a bus workers' strike, there have been traffic problems and John has always arrived late. Today he has an important meeting and his boss is worried that he might not make the meeting in time. He asks a colleague about John and is told not to worry because there is no strike today; therefore, John won't have any problems with the traffic and will get to the meeting on time.

This item challenges us with a deductive task; in particular, propositional reasoning. The formulation of this problem would be as follows:

S: There is a bus strike. L: Arriving late at work

$$S \rightarrow L, \neg S \Rightarrow \neg L$$

This argument reveals a formal fallacy in propositional reasoning: Denying the Antecedent (DA), which bears some similarity with the valid argument known as *Modus Tollens*, or Denying the Consequent (DC). However, as commented above, it is not necessary for the person to answer in these terms, simply to justify his/her answer, indicating for example: "*The reasoning is incorrect because the fact that there is no strike does not imply that John will necessarily not arrive late, because other circumstances could hold him up".*

Regarding the structure of the tests, the items of both are configured around five factors: Deductive Reasoning (DR), Inductive Reasoning (IR) and Practical Reasoning (PR), and Decision making (DM) and Problem Solving (PS). Of all the possible manifestations that include these skills, we chose the most representative structures of each of them owing to their more frequent use in our daily functioning, Each factor comprises 7 items.

Among the problem situations that evaluate Practical Reasoning, four of them measure *argumentation*, which is perhaps the most common form of reasoning (it integrates all the other forms of reasoning), since in our daily activities situations that demand judgements or the production of good arguments to defend points of view, opinions, positions, etc, arise continually. The other three situations evaluate the identification of *fallacies*: errors in reasoning that are misleading, either because of the persuasive force of the argument used or because of the ambiguity of the language employed in most cases prove difficult to identify, and that may prompt us to commit biases in our assessments of arguments. They are frequently used, for example, in the communications media and in politics.

The items that conform the Deductive Reasoning factor evaluate the most important forms of reasoning: *propositional reasoning* (four items) and *categorical reasoning* (3 items). Formal reasoning is less frequent than practical and inductive reasoning, although it is used to a certain extent.

The Inductive Reasoning factor includes (1) causal reasoning, three items; 2) analogical reasoning, two items; 3) hypothetical reasoning, one item and 4) inductive generalisations, one item.

The Decision Making scale assess the use of general procedures of decision, which requires the elaboration of precise judgements about probability and the use of appropriate heuristics in order to make sound decisions Two general situations are included here, in which it is necessary to proceed in a certain way to reach the most appropriate decision, together with the other five, which require the use of the main procedures that we follow when making a decision; namely, heuristics.

Finally, the Problem Solving items, as in the case of Decision Making, are divided into general problems (four items) and specific ones (three items), which are those required to set in motion specific solution strategies.

Both in the Decision Making factor and in that of Problem Solving the use of general procedures of decision and solving are fostered with a view to stimulating the necessary use of strategies for planning a problem. Metacognition and being aware of thinking processes is where actions are planned, directed from, organised and regulated. Again, in this we follow Halpern's proposal.

Below we offer a table summarising the above description of the configuration of the items on the basis of the five factors.

Table 1: PENCRISAL and PENTRASAL factors

DEDUCTION	INDUCTION	PRACTICAL REASONING	DECISION MAKING	PROBLEM SOLVING
PRR=4 CTR=3 **TOT=7**	CR=3 HC=1 AR=2 IG=1 **TOT=7**	ARG=4 FAL=3 **TOT=7**	GRAL=2 PRB=2 CI=1 REP=1 AV=1 **TOT=7**	GRAL=4 RGL=2 ME=1 **TOT=7**

PRR: proposicitional reasoning. CTR: categorical reasoning. TOT: total items. CR: causal reasoning. HC: hipotetical reasoning. AR: analogical reasoning. IG: inductive generalization. ARG: argumentation. FAL: fallacies. GRAL: general decision making. PRB: probabilities judgment. CI: entrapment. REP: representativeness. AV: availability. GRAL: general problem solving. RGL: regularity identification. ME: mean-ends analysis.

The order of presentation of the items is random, although those belonging to the same factor from appearing consecutively are avoided.

The way in which the test is corrected is devised in such a way that it solves the limitation found in the HCTAES answer system, which due to its ambiguities sometimes afforded misleading results. The difficulty is solved by endowing each item with a single solution. In this way, the assessment procedure is simplified and three standard values can be established.

0 points: when the answer given as a solution is **incorrect**

1 point: when the solution is **correct**, but the argumentation is not adequate, showing that the subject only identifies and demonstrates an understanding of the basic concepts

2 points: when as well as giving the correct answer the subject suitably **justifies or accounts** for **why** that decision was reached, thus demonstrating that s/he has made use of more complex processes that involve true mechanisms of production.

Thus, a quantitative scaling system is being used, whose range of values lies between 0 and 71 points as a maximum, for the overall score of the tests and between 0 and 14 for each of the five scales

Example 8: PENTRASAL

Situation 33: Decision making: representativeness

In a study carried out in two secondary education schools, one with 2000 pupils and the other with 200, it is found that the average grades of the boys and girls are very similar. Moreover, in both schools the percentages of boys and girls are more or less the same However, on one test 60% of the girls got better marks than the boys. In which school do you think this happened.

The evaluation of this item, from the real answers, would be as follows:

0 points: when the subjects writes "*the same in both cases, because the probabilities for both are the same regardless of the sample size".* A zero score is given, because the subject does not give the right answer, which would be the smaller school.

1 point: stating "*More probably in the smaller school because there would be fewer girls*". Only one point is given because the subject gives the right answer, but the answer is vague

2 points: for answers such as "*In the smaller one, that of 200 pupils, because the probability would be more real in larger samples than in smaller ones*". Two points are awarded because as well as giving the correct answer the subject has given a suitable justification as to why.

Regarding the time allotted for administration, our tests are defined as power psychometric tests; that is, with no time limit. However, the mean duration estimated for their completion by someone with a normal intellectual level is estimated at between 50 and 60 min. This affords another advantage with respect to the HCTAES open-response test (Halpern, 2003), in which the time required is almost double (120 minutes), and that time (50-60 min) is in fact similar to the time required for application of multiple-choice tests, such as the Cornell Critical Thinking Test (50 minutes) or the California Critical Thinking Skills test of Facione (46 minutes).

The test can be completed in pencil and paper format or with a computer, through the Internet. We have chosen the latter because it offers the most advantages to the corrector, by facilitating the tedious inputting of data, and for the person taking the test, since the programming system allows the test to be taken in several sessions, thereby reducing the possible effects of tiredness that it may cause, especially as regards performance on the last items. The system also allows all the relevant aspects of the test to be controlled, such as preventing any item from not being answered, because the system will not pass to the next item until an answer has been given to the previous one, and preventing the subject from correcting

previous answers or taking the test again once it has been completed. The electronic version also allows both individual and collective application of both tests.

The results finally obtained with the tests are: a global score of the capacity for critical thinking together with the scores for the five subscales (DR, PR, IR, DM and PS), which are those referring to the specific skills of critical thinking.

The preliminary data obtained with these tests point to acceptable psychometric properties, although they can be improved. Currently, we are making the appropriate changes to achieve better quality in the items.

The above description of our assessment proposal offers the improvements that we have posed as a solution to the measurement problem inherent to many of standardised tests. With our method we can be confident that we are activating the thinking mechanisms we are interested in, such that when we eventually assess this activity we can be sure of what we are measuring. This guarantee sheds light on the most important deficiencies in the terrain of the assessment of thinking processes. In order to be able to tackle the complexity of the skills of this class, we must be able to quantify each of them separately. We must measure specific skills with a view to gaining a more precise understanding of what critical thinking is. Our method of assessment seems to offer a reasonable contribution to this conceptual clarification.

As described at the beginning of this work, critical thinking is a multicomponent process. A form of assessment such as that proposed here means that those components can be quantified appropriately. The factors comprising the measurement tests developed by us make the different facets of what critical thinking is crystal clear. It is also reasonably clear that those components are not independent of one another. We can activate some and not others, but in a broad context, such as in our everyday lives, all the processes interact with one another. Thus for example, in the context of practical reasoning we can find all kinds of inferences. The procedure we must follow to make sound decisions also requires us to use our reasoning processes, which allows us to achieve the desired results with greater efficiency. Finally, when we solve a problem we are also making use of our reasoning processes and our capacity to decide soundly. Our assessment method of these skills is able to isolate any of the main thinking processes.

CONCLUSIONS AND SUGGESTIONS

Along this work we have explored the reasons that justify the evaluation of critical thinking; how to address this evaluation, its viability, and the proposal for measurement we offer. Often, the need to assess critical thinking skills stems from an initiative in intervention. When one is seeking to improve these skills, it is necessary to know whether the instruction will produce a change. This is the reason for the assessment. Sometimes, institutional or administrative initiatives aimed at evaluating students' performance in certain intellectual capacities arise. In both cases, the way of estimating these competencies is very similar. For some time, critical thinking assessment programs have been developed for such purposes and with very similar psychometric methods. Tests are constructed that pose problems in which the subject essentially only has to answer by choosing solutions that are already given. This manner of assessing the skills of critical thinking is not very useful because it is not possible to know which skill the person doing the test is using; it is not even possible to be sure that any competency of this type is being used at all. The debate addressed here aims at demonstrating the impossibility of measuring critical thinking with most of the standardised tests published. The way of evaluating such competencies is fraught with severe problems of validity.

From the results of the study we have been carrying out over several years concerning the assessment of critical thinking, we have been able to propose an alternative way of solving this and other important problem. Our approach consists of using task analysis for problem situations. This method has been seen to be efficient when attempting to discern which thinking process is being used in each problem task. In this way, we can know what we are trying to measure and hence diagnose the level of competency achieved by those receiving instruction in critical thinking and design future possibilities for intervention.

Many aspects of our method remain to be improved, but we feel that at least we have established some sound bases that will hopefully allow us to progress with some degree of success. The work to be done involves obtaining good psychometric indices for our tests, since we already have preliminary results that seem to be acceptable.

REFERENCES

Anderson, J. R. (1990). *Cognitive psychology and its implications* (Third edition). San Francisco,CA: W.H. Freeman.

Donders, F. C. (1969). On the speed of mental processes. *Acta Psychologica, 30,* 412-431. Original: 1868.

Ennis, R. H. (2003). Critical thinking assessment. En D. Fasko (Ed.), *Critical thinking and reasoning. Current research, theory, and practice.* (pp. 293-313). Cresskill, NJ: Hampton Press.

Ennis, R. H. (2008). Nationwide testing of critical thinking for higher education: Vigilance required. *Teaching Philosophy, 31*(1), 1-26.

Ennis, R. H., & Weir, E. (1985). *The Ennis-Weir-critical thinking essay test.* Pacific Grove: CA: Midwest Publications.

Govier, T. (1987). *Problems in Argument Analysis and Evaluation.* Dordrecht, Holland: Foris Publications.

Halpern, D. F. (1998). Teaching critical thinking for transfer across domains - Dispositions, skills, structure training, and metacognitive monitoring. *American Psychologist, 53* (4), 449-455.

Halpern, D. F. (2003a). *Thought and knowledge: An introduction to critical thinking* (Fourth edition). Hillsdale, NJ: Erlbaum.

Halpern, D.F. (2003b). The "How" and "Why" of critical thinking assessment. En D. Fasko (Ed.), *Critical thinking and reasoning. Current research, theory, and practice.* (págs. 355-366). New York: Hampton press.

Johnson, R. H. (2008, 8-11 January). *Critical thinking, logic and argumentation.* Paper presented at the Conferencia Internacional: Lógica, Argumentación y Pensamiento Crítico., Santiago de Chile.

Nieto, A.M. y Saiz, C. (2008). Evaluation of Halpern's "Structural Component" for Improving Critical Thinking. *The Spanish Journal of Psychology, 11* (1), 266-274.

Nieto, A.M., Saiz, C. y Orgaz, B. (2009, en prensa). Análisis de la propiedades psicométricas de la versión española del HCTAES-Test de Halpern para la evaluación del pensamiento crítico mediante situaciones cotidianas. *Revista Electrónica de Metodología Aplicada, nº 1* de 2009.

Possin, K. (2008). A field guide to critical-thinking assessment. *Teaching Philosophy, 31* (3), 201-228.

Saiz, C. (2002). Enseñar o aprender a pensar. *Escritos de Psicología, 6,* 53-72.

Saiz, C. y Nieto, A. M. (2002). Pensamiento crítico: capacidades y desarrollo. En C. Saiz (Ed.), *Pensamiento crítico: conceptos básicos y actividades prácticas* (p. 15-19). Madrid: Pirámide.

Saiz, C. y Rivas, S.F. (2008). Intervenir para transferir en pensamiento crítico. *Praxis. 10* (13), 129-149.

Shaw, V. F., & Johnson-Laird, P. N. (1998). Dispelling the "atmosphere" effect on reasoning. En A. C. Quelhas & F. Pereira (Eds.), *Cognition and context.* (pp. 169-199). Lisboa: Instituto Superior de Psicologia Aplicada.