VNiVERSiDAD D SALAMANCA

Departamento de Psicología Básica, Psicobiología y Metodología de las Ciencias del Comportamiento

**FACULTAD DE PSICOLOGÍA**

# *VALIDATION AND PSYCHOMETRIC PROPERTIES OF THE PENCRISAL CRITICAL THINKING TEST*

**Silvia F. Rivas & Carlos Saiz**

**Published in:**

Rivas, S.F and Saiz, C. (2012). Validation and psychometric properties of the PENCRISAL critical thinking test. *Revista Electrónica de Metodología Aplicada. (Electronic Journal of Applied Methodology.)* Vol. 17 (1), 18-34. (http://www.psico.uniovi.es/REMA/v17n1/vol17n1a2.pdf)

**Contact information:**

**Sivia F. Rivas and Carlos Saiz**
Universidad de Salamanca
Facultad de Psicología
Avda. de la Merced, 109-131
37005 Salamanca.
☎ +34 923 29 45 00. Ext. 3278
🗐 +34 923 29 46 08
✉silviaferivas@usal.es; csaiz@usal.es
💻http://www.pensamiento-critico.com

# *VALIDATION AND PSYCHOMETRIC PROPERTIES OF THE PENCRISAL CRITICAL THINKING TEST*

**Silvia F. Rivas & Carlos Saiz**
**Universidad de Salamanca**

**Descriptors:** critical thinking, evaluation, reliability, validity, factor analysis

## Abstract

The purpose of our study was to validate the critical thinking test PENCRISAL in the Spanish population. This test is an appropriate tool to assess reasoning skills (of various kinds, such as argumentation, causal reasoning, analogical reasoning...), problem solving and decision making. The psychometric study was conducted with a sample of 715 Spanish adults, with college cultural level, aged between 18 and 53, and of both sexes. Reliability in terms of internal consistency achieved an acceptable level, if we consider the complexity of the theoretical model of the construct is under critical thinking (Cronbach alpha: ,632). In turn, the reliability in terms of temporal stability, according to the test-retest method, this has proven to be high (r = ,786). And finally, the reliability between judges has reached a high level of agreement between the correctors (Kappa values between ,600 - ,900). Factor analysis has shown a number of factors and subfactors that fit the theoretical model proposed and the results we obtained from the correlations with other tests support the divergent validity, but not convergent. The PENCRISAL is presented as a novel instrument, validated for a Spanish population, whose results show high accuracy and effectiveness as an instrument for measuring the factors that make up the construct of critical thinking.

## 1. INTRODUCTION

There are many ideas on critical thinking, and we must specify which one we advocate. Our thesis is that people reason and make decisions to solve problems or achieve goals. Within this approach we conceive critical thinking as a theory of action. Critical thinking is not merely going more deeply into good judgement and argumentation. It is essential that this reflection prove that critical thinking serves to solve problems or achieve goals, thus considering the argument as a means, not an end. We conceive of critical thinking as an action that leads us to implement our plans.

From this perspective, critical thinking rests on three fundamental skills: reasoning, problem solving and decision making. Thinking must change reality, not just our ideas, it must do more than produce knowledge, it must solve problems. The applied aspect of critical thinking ends in action, effectively solving problems and making sound decisions. Good reflection is essential for this. Therefore, reasoning, deciding and resolving should be considered as inseparable thought mechanisms that are dependent on each other. Good reflection can design a good action plan that is executed with good decision-strategies and problem-solving.

The need and importance of assessing critical thinking in everyday life come from the social or personal desire to improve these skills. To know whether such an improvement exists requires accurate quantification. Thus, the reason for developing the PENCRISAL test stems from the need to evaluate the ARDESOS intervention program we conducted in this area (Saiz and Rivas, 2011) and the lack of suitable instruments. The main difficulties in the assessment of critical thinking are both conceptual and methodological. The former arise from the diversity in conceptualising critical thinking. The latter mostly originate from the fact that the tests for measuring critical thinking (Ennis, 2003) are closed response format instruments, which prevent the exploration of the fundamental mechanisms of thought involved in the task of responding to a test. The HCTAES test (Halpern Critical Thinking Assessment Using Everyday Situations; Halpern, 2006) overcomes this difficulty. This instrument focuses on the thought processes; the items proposed in the test are situations that describe everyday problems that must be resolved through open and closed answers. Our PENCRISAL test stems from this author's approach. We have kept some of her principles, but we have modified some that are not very suitable (Saiz and Rivas, 2008). The principles underlying our test are: 1) the use of everyday situations as items, 2) the use of different domains, in order to assess the degree of generalization of skills, 3) an open answer format that facilitates the exploration of the thought processes, and 4) the use of problem situations with single response that enables the corresponding thought mechanism to be evaluated and facilitates the quantification of the items (see also Rivas and Saiz, 2010).

The aim of this study is the validation of the PENCRISAL Critical Thinking test in the Spanish population.

## 2. METHODOLOGY

### 2.1. Development of the instrument and procedure

In a first phase, an extensive bank of items was prepared in order to have a wide selection available. This first version of the test was applied in a pilot test on a sample of 469 university students from different backgrounds, with the goal of completing the psychometric analysis of the items. From these analyses, we discarded those items that did not provide the psychometric properties needed to remain in the final scale, replacing them with new items and reworking those that could still achieve the properties required.

Based on these results, we developed a second version, which was applied to a sample of 313 university students. Factor analysis on the results of the psychometric study showed a set of factors and sub-factors that explained 59.35% of the total variability. Most of the items (80%) showed they belonged to the expected theoretical factors. In general terms, it can be considered that the scale was already showing good results. This analysis allowed us to make the changes necessary to adjust the properties of the items, primarily to reduce the very high rate of difficulty of some items, and improve the items' fit to theoretical factor model (Saiz and Rivas, 2011).

This study is part of the third phase of the research, which presents the validation of the third version of the test.

### 2.2. Characteristics of the instrument

PENCRISAL is a test consisting of 35 situations-problems that require an open response. The statements are designed so they do not require the response to be

developed and expressed in technical terms but, on the contrary, responses can be easily written in everyday language. These 35 items are configured around five factors: deductive, inductive and practical reasoning, decision making, and problem solving with 7 items per factor (see Annex I). The selection of the most characteristic structures was taken into account in the distribution of situations-problems in each factor. These factors represent the fundamental skills of critical thinking, and within each of them, we find the most relevant ways of reflecting on and resolving issues in our daily activities. The items were presented in random order.

The PENCRISAL test could be done on pencil and paper in group; however we opted to use a computerized test, completed individually over the internet, as this system offered most benefits. These benefits are for the judges, as it facilitates the tedious task of dumping data, as well as for the person performing the test because the system allows the test to be taken in several sessions, the possible effects of fatigue may occur in the person performing the test are reduced, particularly in the performance of the latter items. The system also allows control of all relevant aspects of the test, such as impeding entry of blank items and also provides the opportunity to correct replies or take the test again after completing it. The Internet version allows participants to perform the test anywhere there is a connection to the Net. Other advantages of *online* data collection are well known and we will not dwell on them. Therefore, applying the test over the Internet seems the most suitable way.

The format of the items is open, so that the person must respond to a specific question by adding a justification for his or her answer. For this reason, we have established standardized rating criteria that assign values from 0 to 2 points, depending on the quality of the response:

0 points: when the answer given as a solution to the problem is incorrect;

1 point: when only the solution is correct, but not properly reasoned (identifies and demonstrates understanding of the fundamental concepts);

2 points: when, as well as giving the correct answer, the subject justifies or explains why (where he or she makes use of more complex processes involving real production mechanisms).

In this way, a quantitative scaling system is used; its value range is between 0 and 70 points as the upper limit for the overall test score and 0 – 14 for each of the five scales.

The following is an example of the type of items used in the test:

John needs to use public transport every day to go to work and the journey takes him about two hours. In recent days, with the bus strike, there have been traffic problems, so he has always been late. Today he has a very important meeting and his boss is worried about whether John will arrive in time. John’s boss asks a workmate about John and he tells John's boss not to worry; today there is no strike and so there will not be any traffic problems. Therefore John will arrive in time for the meeting.

**Is John's workmate's conclusion correct? Justify your answer**

As for the time aspect, our test is defined as a psychometric capability test, that is, there is no time limit. The estimated average duration for fully completing the test is 60 to 90 minutes. For more detailed information on the basic principles of the test, see Saiz and Rivas (2008).

The test dimensions should beconsidered multi-dimensionally, in the following terms. Critical thinking as we understand it is related to reasoning and taking decisions to solve issues. These skills should be seen as interrelated. Reaching a goal or solve a problem involves cogitating, making a choice and using good solution strategies. The desired goal is not achieved by using only one of these fundamental activities. The

cooperation of all or part, depending on the situation, is required. For this reason, the dimensions of our test must be understood in the same terms. Deduction and induction, with their different modes, are simply particular forms of reasoning. Reasoning, or explaining, always consists of drawing conclusions from reasons. The difference lies in the way of achieving this. Using analogies or contingency-based relationships requires mechanisms that are sufficiently distinct so as to give meaning to concepts such as analogical or causal reasoning. But the overall purpose is the same in both. This interdependence between the different mechanisms of thought makes the results of our multidimensional validation somewhat difficult to understand. Under this approach, relationships between the dimensions–greater between deduction and induction– and between decision making and problem solving, would be expected. They would also all be interrelated.

### 2.3. Instruments used

**2.3.1. Cornell Critical Thinking Test**(Level Z) (Ennis et al., 1985). This consists of 52 items with three alternative responses. It evaluates the following skills: induction, deduction, observation, credibility, assumptions, and meaning. The test was translated and then conducted over the Internet, maintaining all the original test requirements ($r_{xx}$ between 500 and 770).

2.3.2. **PMA Primary Mental Abilities Test (Thurstone,** 1976). This consists of five basic factors of intelligence: Verbal ($r_{xx}$ =.910), Spacial ($r_{xx}$ = .730), Numeric ($r_{xx}$ = .990), Reasoning ($r_{xx}$ = .920) and Verbal Fluency ($r_{xx}$ = .730).

## 3. SAMPLE

For the final validation of the Spanish version of PENCRISAL, we designed a sample with a minimum size of 784 cases for 95% significance, a power of 80%, p = q = 0.50 and a maximum error of 3.5 %. It was decided to use a purposive sampling method and for convenience, given the logistical impossibility of finding subjects by srm. (simple random sampling). The sample finally achieved was similar in size, 753 cases, although some were eliminatedduring previous exploratory analysis owing to incomplete questionnaires, malicious responses denoting lack of participation, and outliers. Finally, the total number of cases analysed was 715 (91.1% of the target), providing a perfect representation of the adult Spanish population at university cultural level.

Of these 715 participants, 30.8% (220) were men and 69.2% (495) were women. The mean age of the sample was 24.35 years (95% CI: 23.88 to 24.81) with a standard deviation of 6.28 years. This variable is not normally distributed with p <.050 (KS Test: Z = 5.89, p = .000) due to a marked positive asymmetry (As = 1.502) and at greater values than is normal (K = 2.33). With a median of 21, the 50% mid-range was between 20 and 28 years of age.The full range was 18–53 years. The mean age of the men was 24.90 (95% CI: 24.03 to 25.76), and of the women, 24.10 (95% CI: 23.56 to 24.64).This difference is not significant with p>.050 (Student's t-distribution: t = 1.56, df 713, p =.118). By ranges, 57.5% (411 cases) were still of university degree age (to 22 years), and 42.5% (304) were university post-graduates or in employment.

This sample of 715 cases was used for item analysis, internal consistency, factorial validation and descriptive study, and the construction of the scale. For temporal stability studies, interrater reliability and convergent-discriminant validity,

different subsamples, drawn randomly from the 753 initial participants were used before starting with the statistical analyses, in an attempt to avoid bias.

## 4. STATISTICAL ANALYSIS

Data analysis was performed using the statistical package SPSS-IBM-19. We used the Kolmogorov-Smirnov (KS) goodness of fit test to verify that the numerical variables were modelled on the Gaussian bell. The item analysis was performed using the difficulty index and homogeneity index corrected between the item and the total score on the scale, estimated using Pearson. For reliability analysis we usedCronbach's*alpha* and Pearson for temporal stability. The interrater reliability coefficients were found with Cohen's Kappa for each of the items. Construct validity was analysed using Principal Components Analysis, testing out different methods of rotation, both orthogonal and oblique. After comparing their solutions and similarity, we finally decided to opt for those found through the Varimax method. We had previously tested the factorization conditions using the Bartlett and Kaiser-Meier-Olkin tests together with determinant of the correlation matrix. Correlations for convergent and divergent validity were performed using Pearson coefficients.

## 5. RESULTS

### 5.1. Pencrisal Scores and Scaling

Total Pencrisalscores in the sample of 715 participants analysed, are distributed with a mean 27.48 (95% CI: 27.00 to 27.95) and 6.49 standard deviation for a range of scores: 12-44. The distribution of these values shows very slight deviation from the normal model of the bell curve with p <, 050 but tolerable (p =, 039>, 001 in the KS test). We built a scale in percentiles for the general population, given that there are no significant differences by gender or age, and each of the factors (see Table 1).

**Table 1.**

PENCRISAL: Scales for general population

| Centiles | Raw Scores | | | | | |
|---|---|---|---|---|---|---|
| | DR | IR | PR | PS | DM | Tot. |
| 99 | 10 | 9 | 12 | 11 | 10 | 41 |
| 95 | 8 | 8 | 10 | 10 | 9 | 38 |
| 90 | 7 | 7 | 9 | 9 | 9 | 36 |
| 85 | 7 | 7 | 9 | 9 | 8 | 35 |
| 80 | 6 | 6 | 8 | 8 | 8 | 33 |
| 75 | 6 | 6 | 8 | 8 | 8 | 32 |
| 70 | 5 | 6 | 7 | 7 | 7 | 31 |
| 65 | 5 | 6 | 7 | 7 | 7 | 30 |
| 60 | 5 | 5 | 6 | 7 | 7 | 29 |
| 55 | 4 | 5 | 6 | 7 | 7 | 29 |
| 50 | 4 | 5 | 5 | 6 | 6 | 28 |
| 45 | 4 | 5 | 5 | 6 | 6 | 27 |
| 40 | 4 | 5 | 5 | 5 | 6 | 26 |
| 35 | 3 | 4 | 4 | 5 | 5 | 25 |
| 30 | 3 | 4 | 4 | 5 | 5 | 24 |
| 25 | 3 | 4 | 4 | 4 | 5 | 23 |
| 20 | 3 | 4 | 3 | 4 | 5 | 22 |
| 15 | 2 | 3 | 3 | 3 | 4 | 20 |
| 10 | 2 | 3 | 3 | 3 | 4 | 18 |
| 5 | 1 | 2 | 2 | 2 | 3 | 16 |
| 1 | 0 | 1 | 1 | 1 | 2 | 13 |
| N | 715 | 715 | 715 | 715 | 715 | 715 |
| Mean | 4.42 | 5.03 | 5.78 | 6.04 | 6.21 | 27.48 |
| σ | 2.16 | 1.63 | 2.58 | 2.39 | 1.91 | 6.49 |

## 5.2. Item Analysis

The PENCRISAL is configured as a difficult test to complete. This is necessary in this type of testing because only in this way can we show the effect of the intervention, without having to design another parallel instrument for this purpose. After the above, the difficulty of the items ranges from 0.80 to 0.06 with a mean of 0.39 (95% CI: 0.34 to 0.45) and 0.16 standard deviation. Of these, 18 items (51.4%) are in the medium difficulty range, 3 (8.6%) are easy (ID> 0.65) and the remaining 14 (40%) can be considered as very difficult (ID <0.35).

Each scale's corrected homogeneity index, with respect to the total, is highly significant in all with p <.001. The range of these indices is: .172 - .383.

## 5.3. Internal Consistency and Reliability

The reliability study was carried out from the perspective of internal consistency, temporal stability and interrater consistency, the latter being a crucial question, given the peculiarities of the method of correcting the test. The internal consistency of the 35 items was estimated by Cronbach's Alpha method. The reliability coefficient obtained is .632 highly significant with p <.001 (n = 715; Anova: F=174.73; 34 y 24276 df; p=.000), indicating that the degree of homogeneity among items is quite acceptable.

The reliability and temporal stability was estimated by the retest method. We selected a random subsample of 130 subjects who took the test again between 4 and 5 weeks after the first one. The results show good stability with high and significant Pearson coefficients both in the total score (r =.786, p <.001) and for each of the subscales. See table 2.

**Table 2.**
Reliability according to the retest method

| Variables | 1st application | | Retest | | Correlation test-retest | |
|---|---|---|---|---|---|---|
| | Mean. | σ | Mean. | σ | r | p |
| TOTAL. | 26.44 | 5.49 | 26.61 | 5.31 | .786 | .000 |
| D.R. | 3.93 | 1.93 | 3.83 | 1.95 | .599 | .000 |
| I.R. | 5.12 | 1.34 | 5.18 | 1.63 | .467 | .000 |
| P.R. | 5.52 | 2.07 | 5.79 | 1.92 | .465 | .000 |
| D.M | 5.94 | 1.87 | 5.78 | 1.97 | .548 | .000 |
| P.S. | 5.93 | 2.08 | 6.04 | 2.11 | .556 | .000 |

For interrater reliability, given the complexity that the correction of the test items requires, we selected another random subsample of 100 participants. These questionnaires were corrected independently by three judges properly trained in this task. During this process some incomplete questionnaires were observed, so that the number of cases analysed for this part of the study, varies between 91 and 96. Subsequently the results of the 3 judges were crossmatched with each other and all Cohen Kappa coefficients were estimated. The results can be seen in Table 3 and they show that in all cases coefficients greater than 0.500 were found, most with values above 0.600 and so they can therefore be classified as good concordance according to the Landis and Koch criteria (1977). The mean correlation between judges 1 and 2 is 0.738 (95% CI: 0.70 to 0.78) with a range of .515 to .970. The mean correlation between judges 1 and 3 is 677 (95% CI: 0.64 to 0.72) with a range of .510 to .979. And

finally, the mean correlation between judges 2 and 3 is 0.627 (95% CI: 0.59 to 0.66) with a range of .503 to .939. All these indices have proved highly significant with p <.001.

**Table 3.**
Interrater Kappa values

| Deduction | C1-C2 | C1-C3 | C2-C3 |
|---|---|---|---|
| 1 | .727 | .586 | .594 |
| 3 | .970 | .587 | .564 |
| 5 | .716 | .657 | .546 |
| 8 | .637 | .606 | .503 |
| 16 | .827 | .821 | .664 |
| 23 | .834 | .539 | .535 |
| 28 | .862 | .597 | .666 |
| Induction | C1-C2 | C1-C3 | C2-C3 |
| 2 | .553 | .662 | .519 |
| 4 | .919 | .838 | .738 |
| 6 | .630 | .769 | .572 |
| 9 | .659 | .622 | .556 |
| 10 | .608 | .628 | .657 |
| 24 | .565 | .510 | .552 |
| 29 | .658 | .590 | .580 |
| P.R | C1-C2 | C1-C3 | C2-C3 |
| 7 | .667 | .716 | .547 |
| 11 | .752 | .637 | .606 |
| 21 | .674 | .647 | .646 |
| 25 | .630 | .758 | .677 |
| 30 | .760 | .828 | .711 |
| 31 | .718 | .598 | .569 |
| 34 | .908 | .536 | .519 |
| D.M | C1-C2 | C1-C3 | C2-C3 |
| 14 | .785 | .692 | .581 |
| 17 | .721 | .827 | .605 |
| 18 | .844 | .663 | .752 |
| 19 | .515 | .670 | .540 |
| 20 | .672 | .558 | .601 |
| 27 | .742 | .643 | .609 |
| 32 | .699 | .665 | .661 |
| P.S | C1-C2 | C1-C3 | C2-C3 |
| 12 | .835 | .949 | .879 |
| 13 | .747 | .590 | .661 |
| 15 | .959 | .979 | .939 |
| 22 | .733 | .632 | .522 |
| 26 | .729 | .516 | .615 |
| 33 | .858 | .903 | .901 |
| 35 | .717 | .665 | .544 |
| Mean K indices | .738 | .677 | .626 |
| P | <000 | <.000 | <.000 |

## 5.4. Construct validity

The different levels of our mental activity must be related and integrated to be effective in action. Thus, given the demonstrated multidimensionality of the critical thinking construct, it was decided to begin the study of the construct validity applying factor analysis independently to each of them. In all these analysis the prerequisites of sampling adequacy (KMO>, 500) and sphericity (Bartlett's test with p <.001) have been satisfactorily completed, with determinants of correlation matrices close to zero. The specific values in each case are in the respective tables, where it can be seen that in all cases the conditions necessary for the use of this statistical technique are fulfilled.

Below are the results for each of the five dimensions, which show the proper adjustment to the initial theoretical model, which had appeared in studies using older versions of the test.

a) *Deduction* See table 4. It can be seen that four items are grouped around the *Propositional* deduction factor with a loading in the range of .495 – .720. The other three items are grouped around the*Categorical* inference subfactor with factor loadings between .597 – .706. The total internal variability explained by

the items of this dimension is 44.83%. The deduction dimension explains almost 10% of the total Pencrisalvariability.

**Table 4.**
Factor structure and reliability of the scale: DEDUCTION (7 items)
*Conditions:* KMO= .634; Bartlett p<.001

| Component | No. items | Items | αCronbach | Factors | % Of variance explained in its Dimension | % of Total variance explained |
|---|---|---|---|---|---|---|
| PropositionalDed. | 4 | 1; 3; 8; 16 | - | 1 | 25.43 | 5.48 |
| Categorical Ded. | 3 | 5; 23; 28 | - | 1 | 19.41 | 4.18 |
| Total Dimension | 7 | - | .371 | 2 | 44.83 | 9.66 |

b) *Induction*. See table 5. Three items with factor loadings within the .562 – .674 range are configured as *analogical* reasoning. Two further define the *Causal* inductive factor, with loadings of .649 and .816. And the last two, with loadings of .680 and .765 are *Verification* procedures (hypothesis testing (IT) and inductive generalizations). The internal variability explained by all of them reaches 50.59%, while the induction factor explains almost 11% of the total variability of the test.

**Table 5.**
Factor structure and reliability of the scale: INDUCTION (7 items)
*Conditions:* KMO= .575; Bartlett p<.001

| Component | No. items | Items | αCronbach | Factors | % Of variance explained in its Dimension | % of Total variance explained |
|---|---|---|---|---|---|---|
| Induct. Reason. Analogue. | 3 | 6; 9; 24 | - | 1 | 19.23 | 4.14 |
| Induct. Causal | 2 | 2; 10 | - | 1 | 16.02 | 3.46 |
| Induct. Proc. Verification (HT and IG) | 2 | 4; 29 | - | 1 | 15.32 | 3.30 |
| Total Dimension | 7 | - | .250 | 3 | 50.59 | 10.90 |

c) *Practical reasoning*. See table 6. Four items are grouped in *the Argumentation* dimension, with factor loadings included in the .525 – .753 range, while the other three make up the Fallacies component with loadings ranging from .483 –to .634. The total variability accounted for internally reaches 40.38%. The practical reasoning dimension explains about 9% of the total variability.

Table 6.
Factor structure and reliability of the scale: Practical reasoning (7 items)
*Conditions*: KMO= .624; Bartlett p<.001

| Component | No. items | Items | αCronbach | Factors | % Of variance explained in its Dimension | % of Total variance explained |
|---|---|---|---|---|---|---|
| Argumentation | 4 | 7; 21; 25; 30 | - | 1 | 24.05 | 5.18 |
| Pract.Reas.: Fallacies | 3 | 11; 31; 34 | - | 1 | 16.32 | 3.52 |
| Total Dimension | 7 | - | .425 | 2 | 40.38 | 8.70 |

d) *Decision making.* See table 7. In this component foursubfactors were identified, all constituting 2 items, since one of them loads two of the sub-factors identified. The *General DM* factor with loading exceeding .806 explains the 19.70 internal variability factor. The *Probability DM*, with weights of .512 and .859, explains the 15.02%. The *General Heuristic DM*(representativeness and availability) with saturations of .523 and .698 explains the 17.21%. And finally, the *DM specific heuristics* (availability and cost of investment), with saturations of .527 and .905

explains the 15.94%. As can be seen, the availability item loads these two sub-factors. This is understood in general heuristics because, although the two are conceptually different items, they both initiate the same type of general strategies for estimating probabilities of events. However, the investment cost of the specific heuristic DM factor is a strategy that depends in part on availability and not on representativeness. For this reason, it is grouped as a different factor from the general one. The total internal variability explained reaches 67.87%. The decision-making component is the one with greatest weight in the entire test because it accounts for 14.61% of the total variability.

**Table 7.**
Factor structure and reliability of the scale: DECISION MAKING (7 items)
*Conditions:* KMO= .575; Bartlett p<.001

| Component | No. items | Items | αCronbach | Factors | % Of variance explained in its Dimension | % of Total variance explained |
|---|---|---|---|---|---|---|
| General DM | 2 | 14; 27 | - | 1 | 19.70 | 4.24 |
| DM: General heuristics (Rep and AV) | 2 | 19; 20 | - | 1 | 17.21 | 3.71 |
| DM: Specific heuristics (AV and CI) | 2 | 18; 20 | | 1 | 15.94 | 3.43 |
| Probability DM | 2 | 17; 32 | - | 1 | 15.02 | 3.23 |
| Total Dimension | 7 | - | .213 | 4 | 67.87 | 14.61 |

e) *Problem solving* (PS). See table 8. Four items have been framed in the *general P.S.* subfactor with value loadings in the range .511 – .710, while the other three items are the *specific PS*component (search for regularities and means-end analysis) with factor loadings between .548 – .705. These items account for 38.96% of the total specific variability; the P.S. factor accounts for 8.4% of the total Pencrisalvariability.

**Table 8.**
Factor structure and reliability of the scale: PROBLEM SOLVING (7 items)
*Conditions:* KMO= .624; Bartlett p<.001

| Component | No. items | Items | αCronbach | Factors | % Of variance explained in its Dimension | % of Total variance explained |
|---|---|---|---|---|---|---|
| General P.S. | 4 | 13; 22; 26; 35 | - | 1 | 19.56 | 4.21 |
| Specific P.S (RGL and ME) | 3 | 12; 15; 33 | - | 1 | 19.40 | 4.18 |
| Total Dimension | 7 | - | .373 | 2 | 38.96 | 8.39 |

We calculated the correlations between the five factors described above and with the total score (see Table 9). Correlation coefficients statistically significant are obtained given the sample size, but with intensities between factors (from .103 to .291). This supports the multidimensionality of the construct and independence between factors.

**Table 9.**
Matrix of intercorrelations of the factors with the PENCRISAL total

| | | DR | IR | PR | PS | DM | Total |
|---|---|---|---|---|---|---|---|
| DR | Pearson Correlation | ____ | | | | | |
| | Sig | | | | | | |
| | N | | | | | | |
| IR | Pearson Correlation | .204 | ____ | | | | |
| | Sig | .000 | | | | | |
| | N | 715 | | | | | |
| PR | Pearson Correlation | .254 | .289 | ____ | | | |
| | Sig | .000 | .000 | | | | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | N | 715 | 715 | | | | |
| PS | Pearson Correlation | .103 | .235 | .291 | ____ | | |
| | Sig | .003 | .000 | .000 | | | |
| | N | 715 | 715 | 715 | | | |
| DM | Pearson Correlation | .115 | .149 | .176 | .206 | ____ | |
| | Sig | .001 | .000 | .000 | .000 | | |
| | N | 715 | 715 | 715 | 715 | | |
| Total | Pearson Correlation | .558 | .569 | .713 | .638 | .516 | |
| | Sig | .000 | .000 | .000 | .000 | .000 | ____ |
| | N | 715 | 715 | 715 | 715 | 715 | |

As for the factor analysis of the full set of 35 items (KMO = .683; Bartlett test: $Chi^2$=1988.39: 595 df, P = .000) reveals the existence of factors and sub-factors of 13 components that match the previous breakdown: 2 in deduction, 3 in induction, 2 in practical reasoning, 2 in problem solving and the remaining 4 in decision-making. The loadings of the items are in the .400 – .762 range. The total test variability explained by this set of factors and sub-factors approaches 53% as shown in Table 10.

**Table 10.**
Variability explained in the AF of CP with Varimax rotation of the entire test (35 items)

| | Sum of squared loadings of the rotation | | |
|---|---|---|---|
| Component | Total | % of the variance | % accumulated |
| 1 | 1.914 | 5.467 | 5.467 |
| 2 | 1.692 | 4.835 | 10.303 |
| 3 | 1.664 | 4.755 | 15.057 |
| 4 | 1.469 | 4.198 | 19.255 |
| 5 | 1.464 | 4.184 | 23.439 |
| 6 | 1.396 | 3.988 | 27.427 |
| 7 | 1.369 | 3.911 | 31.338 |
| 8 | 1.299 | 3.713 | 35.051 |
| 9 | 1.281 | 3.661 | 38.712 |
| 10 | 1.277 | 3.647 | 42.359 |
| 11 | 1.275 | 3.642 | 46.001 |
| 12 | 1.201 | 3.433 | 49.434 |
| 13 | 1.153 | 3.293 | 52.727 |

## 5.5. Convergent and divergent validity

For analysis of both types of validation we took a new random subsample of 130 participants. In the previous exploratory study it was decided to eliminate any cases with outliers, but the loss was minimal.

For this part of the study, the Cornell critical thinking test was applied, because this is one of the most widely used.

After checking the linearity of the relationship, we proceeded to correlate Pencrisalscores with the Cornell scores (see Table 11). The coefficients obtained are mostly not statistically significant (p>.050). These results do not support the convergent validity

**Table 11.**
Correlations between Cornell and PENCRISAL

| | | DR PENCRISAL | IR PENCRISAL | PR PENCRISAL | PS PENCRISAL | DM PENCRISAL | TOTAL PENCRISAL |
|---|---|---|---|---|---|---|---|
| IR CORNELL | Pearson Correlation | .080 | .101 | .128 | .192 | .130 | .211 |
| | Sig | .372 | .258 | .150 | .031 | .146 | .017 |
| | N | 127 | 127 | 127 | 127 | 127 | 127 |
| DR CORNELL | Pearson Correlation | .099 | .125 | .000 | -,083 | .092 | .059 |
| | Sig | .269 | .161 | .998 | .355 | .301 | .513 |
| | N | 127 | 127 | 127 | 127 | 127 | 127 |
| TOTAL CORNELL | Pearson Correlation | .066 | .220 | .197 | .152 | .046 | .224 |
| | Sig | .461 | .013 | .026 | .088 | .611 | .001 |
| | N | 127 | 127 | 127 | 127 | 127 | 127 |

For divergent validity, the PMA Intelligence test was applied (Primary Mental Abilities). The correlations are mostly not significant (p> .050) and those that were showed low levels of intensity (r <.200) which clearly demonstrates the absence of theoretical association among tests, and defends divergence (table 12).

**Table 12.**
Correlations between Cornell and PENCRISAL

| | | DR PENCRISAL | IR PENCRISAL | PR PENCRISAL | PS PENCRISAL | DM PENCRISAL | TOTAL PENCRISAL |
|---|---|---|---|---|---|---|---|
| PMA.V | Pearson Correlation | -,067 | .165 | .114 | .198 | .109 | .169 |
| | Sig | .454 | .063 | .202 | .025 | .221 | .057 |
| | N | 127 | 127 | 127 | 127 | 127 | 127 |
| PMA.S | Pearson Correlation | .072 | -.010 | .140 | .141 | .157 | .174 |
| | Sig | .418 | .910 | .117 | .114 | .077 | .050 |
| | N | 127 | 127 | 127 | 127 | 127 | 127 |
| PMA.R | Pearson Correlation | .025 | .109 | .001 | .199 | .204 | .169 |
| | Sig | .778 | .221 | .992 | .025 | .021 | .057 |
| | N | 127 | 127 | 127 | 127 | 127 | 127 |
| PMA.N | Pearson Correlation | -.093 | .031 | -.002 | -.028 | -.020 | -.040 |
| | Sig | .298 | .733 | .987 | .754 | .824 | .652 |
| | N | 127 | 127 | 127 | 127 | 127 | 127 |
| PMA.F | Pearson Correlation | .157 | -.058 | .137 | -.120 | -.033 | .037 |
| | Sig | .078 | .519 | .124 | .180 | .716 | .682 |
| | N | 127 | 127 | 127 | 127 | 127 | 127 |
| PMA TOTAL | Pearson Correlation | .057 | .049 | .159 | .126 | .143 | .181 |
| | Sig | .525 | .583 | .074 | .157 | .110 | .041 |
| | N | 127 | 127 | 127 | 127 | 127 | 127 |

## 6. CONCLUSIONS

PENCRISAL is a useful and innovative tool for assessing critical thinking skills, and has proven its validity in the Spanish population of university level education.

PENCRISAL provides a number of advantages in evaluation: 1) This very innovative measurement tool, together with the HCTAES, are the only tests of critical thinking focused on the processes of critical thinking, 2) it contributes to improving the assessment of critical thinking skills in an integrated manner, as there are currently no instruments of this nature in Spanish, and 3) using everyday situations that can be resolved in only one way as items and an open response format, makes PENCRISAL an accurate tool for measuring critical thinking.

Regarding the study of the psychometric properties of the test has we have demonstrated statistically the suitable factor structure adjustment of the test to the proposed theoretical model outlined in adult Spanish population of university cultural level. Also, regarding convergent and divergent validity, the PENCRISAL test has shown high divergent power in comparison with theoretical intellectual capacity constructs. In turn, the absence of other specific tests that measure the same traits and

the same kind of open response format makes it difficult to achieve a strong convergent validity. The absence of validity with respect to Cornell is due to the nature of the instruments. Cornell is a closed comprehension test while PENCRISAL, conversely, is open and productive. This means the way of responding is sufficiently different to produce different results. What is required in each one makes the difference. Our test requires developing an explanation for each answer; Cornell only requires a choice. So the performance and nature make obtaining such validity difficult. However, this highlights the differential and innovative nature of this test with respect to those currently existing in the field of critical thinking skills assessment.

As for the reliability study, high temporal stability has been proven through the measuring instrument retest procedure.

Finally, one of the most important aspects of the instrument is the interrater reliability study, since, given the special characteristics of the type of test, the correction system requires a necessarily high degree of agreement among the correctors. It has been possible to demonstrate a high level of concordance with each of the three evaluators. These correlations, both in total score on the test and in the five factors, are highly significant and show high correlation values.

Among the limitations of the test we can highlight, first, that the construct that evaluates critical thinking skills is a complex construct that can be defined from many different theoretical frameworks, resulting in different kinds of instruments. Second, PENCRISAL has the limitations of all open answer tests. The correction system requires expert evaluators, and the time required for correcting response protocols is high. Finally, we are aware that the factorial analysis procedure by dimensions is not very common but it has been used because of the singularity and complexity of the test. As we have seen, the composition of factor analysis together with all of the 35 items corresponds exactly to each of the factors described in subfactors and analysis by dimensions. It would be much more complex for the reader to interpret and understand from the matrix of the 13 components, than from each separate factor. That is why it is presented in this way. It is clear that, seen in their entirety, 35 items grouped into 13 factors and subfactors involve 2 or 3 items per factor, which is not ideal. But given the time required to take the test it is not advisable to add more items; the instrument would not practical in terms of the time required because respondents must justify their replies, that is, they must produce an extended response. From these considerations, for the future we propose to convert the test into a battery composed of five subscales that correspond to the five studied theoretical constructs; thus the test could have a greater number of items for each one of them.

Given the characteristics of the PENCRISAL test, we consider it could be widely applied, covering educational, social, personal and research areas, and it could also be an appropriate tool for evaluating the effectiveness of instructional programs and for improving critical thinking skills. However, for the future, we must further improve some aspects of the test that are determined by the limitations noted above. It is important to work on the instrument in order to achieve greater dimensional accuracy by merging some of the subfactors proposed. Also, given the complexity and nature of the test, and because we are aware that the psychometric indices could be improved, it would be appropriate to make a greater effort in this direction. And finally, it would be necessary to develop an automated test correction system using semantic categorization procedures. All these improvements have been implemented in the several projects we are developing.

## 7. REFERENCES


Ennis, RH (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., Millman, J., &Tomko, T.N. (1985). *Cornell Critical Thinking Test, Level X & Level Z-Manual* (3rd ed.). PacificGrove, CA: Midwest.

Halpern, D.F. (2006). *Halpern Critical Thinking Assessment Using Everyday Situations: Background and scoring standards*. Unpublished report.

Rivas, SF and Saiz, C. (2010). ¿Esposibleevaluar la capacidad de pensarcríticamente en la vidacotidiana?En Jales, H.R. y Neves, J. (Eds.), *O Lugar da Lógica e da Argumentação no Ensino da Filosofia* (53-74). Coimbra: Unidade I&D, Linguagem, Interpretação e Filosofia

Saiz, C. y Rivas, S.F. (2008).Evaluación en pensamientocrítico: unapropuestaparadiferenciarformas de pensar*Ergo, Nueva Época, 22-23,* 25-66.

Saiz, C. y Rivas, S.F. (2011).Evaluation of the ARDESOS program: an initiative to improve critical thinking skills. *Journal of the Scholarship of Teaching and Learning, Vol. 11, No. 2,* 34-51.

Thurstone, L.L., &Thurstone.T.G. (1976).*PMA: Aptitudes MentalesPrimarias (*Primary Mental Abilities in English). Madrid: TEA

# APPENDIX I

**DISTRIBUTION OF PENCRISAL ITEMS AND FACTORS**

| | FACTORS | | | | |
|---|---|---|---|---|---|
| **Item** | **DEDUCTION** | **INDUCTION** | **PRACTICAL REAS.** | **DECISION MAKING** | **PROBLEM SOLVING** |
| **1** | Propositional R. | | | | |
| **2** | | Causal R. | | | |
| **3** | Propositional R. | | | | |
| **4** | | HypotheticalR. | | | |
| **5** | Categorical R. | | | | |
| **6** | | Causal R. | | | |
| **7** | | | Argumentation | | |
| **8** | Propositional R. | | | | |
| **9** | | Analogical R. | | | |
| **10** | | Causal R. | | | |
| **11** | | | Fallacy | | |
| **12** | | | | | Regularity identification |
| **13** | | | | | General |
| **14** | | | | General | |
| **15** | | | | | Regularity identification |
| **16** | Propositional R. | | | | |
| **17** | | | | Probability. | |
| **18** | | | | Investment cost | |
| **19** | | | | Representativeness | |
| **20** | | | | Availability | |
| **21** | | | Argumentation | | |
| **22** | | | | | General |
| **23** | Categorical R. | | | | |
| **24** | | Analogical R. | | | |
| **25** | | | Argumentation | | |
| **26** | | | | | General |
| **27** | | | | General | |
| **28** | Categorical R. | | | | |
| **29** | | Inductive Generalization | | | |
| **30** | | | Argumentation | | |
| **31** | | | Fallacy | | |
| **32** | | | | Probability. | |
| **33** | | | | | Means End |
| **34** | | | Fallacy | | |
| **35** | | | | | General |

| **DEDUCTION 7 items** | **Induction 7 items** | **Practical Reas. 7 items** | **Decision Making 7 items** | **Problem Solving 7 items** |
|---|---|---|---|---|
| **PR.R = 4<br>CT.R = 3** | **CR = 3<br>HC = 1<br>AR = 2<br>IG = 1** | **ARG = 4<br>FAL = 3** | **GRAL = 2<br>PRB = 2<br>IC = 1<br>REP = 1<br>AV = 1** | **GRAL = 4<br>RGL = 2<br>ME = 1** |