# *Multiple-Choice Tests with Correction Allowed in Autism: An Excel Applet*

**Professor Elisabetta Monari Martinez**

Department of Pure and Applied Mathematics,
University of Padua, Padua, Italy

## *Abstract*

The valuation of academic achievements in students with severe language impairment is problematic if they also have difficulties in sustaining attention and in praxic skills. In severe autism all of these difficulties may occur together. Multiple-choice tests offer the advantage that simple praxic skills are required, allowing the tasks to be performed without physical support. Even so, attentive and behavioral difficulties may be so disruptive that achievements may be underestimated. Since special needs educators can give immediate feedback on each answer, a strategy might be to permit corrections, allowing further attempts, in order to mitigate these problems and to better capture their knowledge. Here a Microsoft Excel applet is designed to compute the statistical significance and the final grade of multiple-choice tests, if up to two corrections per selection are allowed. The method was used with a nonverbal student with severe autism and Down syndrome in a mainstream secondary school.

## *Multiple-Choice Tests with Correction Allowed in Autism: An Excel Applet*

A conventional multiple-choice (MC) test item consists of a stem (the question) and a list of alternatives (possible answers to the question). The stem may be also an incomplete statement and the alternatives its possible completions. Exactly one alternative is the *correct answer* and the others are *distracters*. Possible weaknesses in a MC test are that it does not measure what it is supposed to, that it contains clues to the correct answer and that it is worded ambiguously (Burton, S.J., Sudweeks, R.R., Merrill, P.F., Wood, B., 1991). A great deal of research on the use of multiple-choice tests in education has been conducted (for a review, see Haladyna, Downing, Rodriguez, 2002). That research has inspired a number of well-known guidelines that describe how to prepare the items for an MC test, taking account of the influence that item format exerts on students' comprehension and outcomes (e.g. Martinez, 1999).

However, psychometric research also reveals the problems that can emerge concerning the reliability and validity of MC test results, highlighting the role that guessing can play in test outcomes alongside a suite of other factors, unrelated to the aim of the test, that can influence students' choices. For instance, when students are not certain of the correct answer, their choices may be influenced by the apparent likelihood of the alternative answers. Since 1919, *formula-scoring* procedures have been used to mitigate these problems (Thurstone 1919). The most popular procedures used at present, are: (S1) simply to compute the percentage of the right answers with respect to number of questions and (S2) give 1 point for each right answer, no point for the unanswered questions, and a penalty of -1/(*c*-1) points for each failure, where *c* is the fixed number of alternatives per question. In this latter case, the grade will be the percent ratio of the sum of the points to the number of questions. With the S1 procedure, the test is sensible to guess when the answer is uncertain, because omissions and failures are computed in the same way (zero score), while, with the S2 procedure, the penalty for errors depends on the probability, 1/*c*, of guessing correctly in response to each question, as we will see later. For instance, S1 is used in the American College Testing (ACT) exam and S2 in the Scholastic Aptitude Test (SAT) exam. Some studies support the first, while others (for different reasons) support the second; given a large number of questions, both approaches yield equivalent reliability (Prieto & Delgado, 1999).

We employ S2 in the material that follows, because it appears to be more reliable for tests involving small numbers of questions (figure 1).

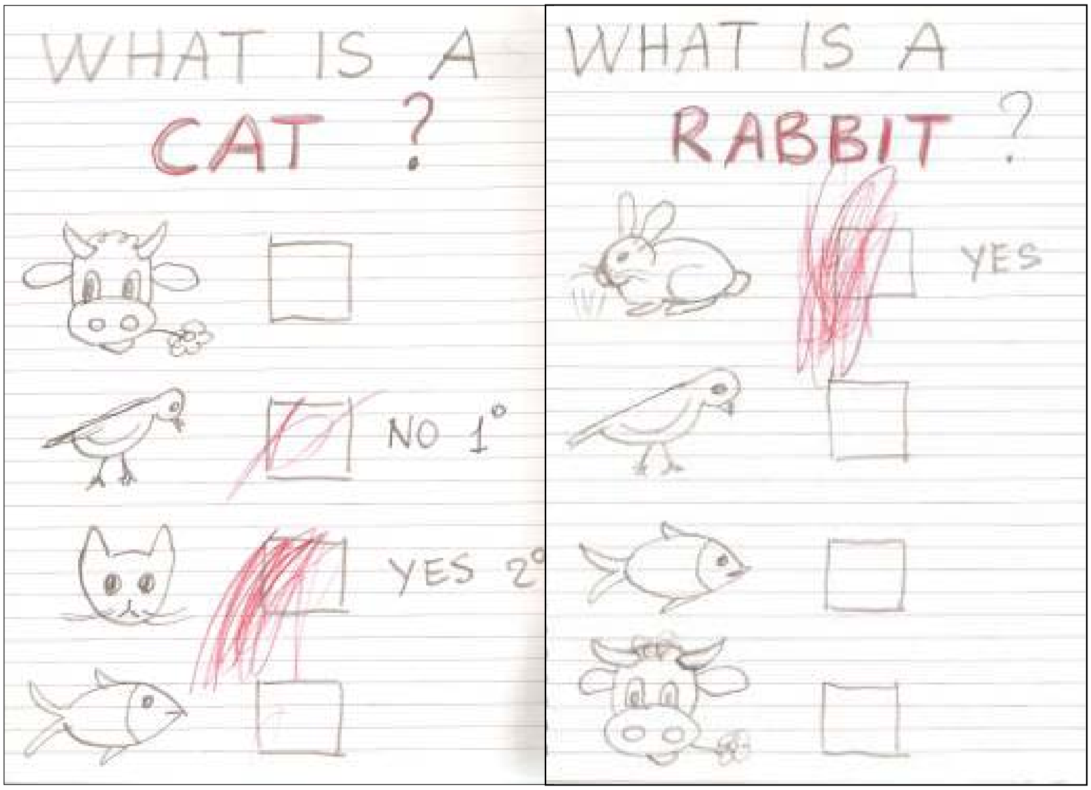
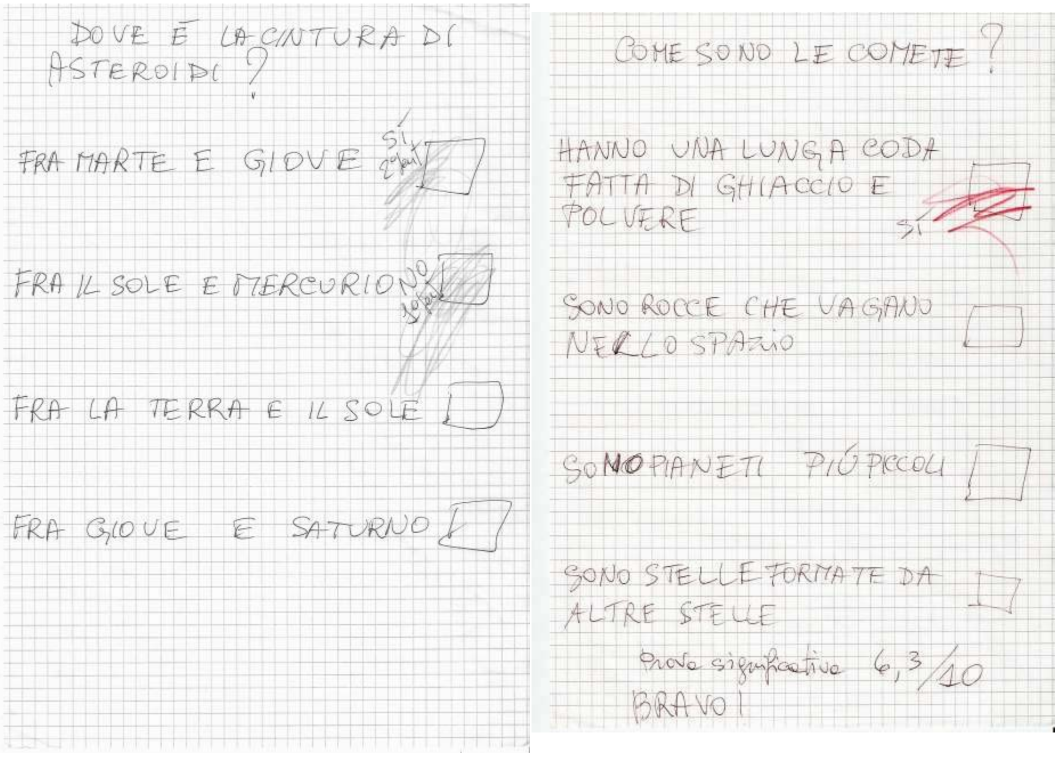


Fig. 1

**Figure 1**. *These are some examples of how to answer the test questions, when some corrections are allowed. To avoid influencing the choice, no physical support is permitted. In the first page, on an A4 sheet, the first answer was wrong - perhaps because that option is spatially close to the correct one – and the second was correct. In the second example, the correct answer was also the first selected, and the selection was relatively certain, as indicated by the breadth and intensity of the student's mark. These examples are taken from a test conducted by a 16-year-old boy with autism with Down syndrome, who studied English as second language.*

Multiple-choice tests are frequently used in special education, to test the performance of nonverbal students, who are difficult to test in other ways. Often, they are employed as components of popular tools like the Peabody test (PPVT™-4: Peabody Picture Vocabulary Test, 2006), used to measure students' vocabulary and word comprehension. There are also studies that describe the influence of particular picture types, employed as alternatives, on performance outcomes (e.g. Heuer & Hallowell, 2007). Here, we propose the use of MC tests to evaluate the academic achievements of students with autism (Twachtman-Cullen, D., 2006; Volkmar, Paul, Klin, Cohen, 2005; Schopler & Mesibov, 1995), when students exhibit poor or absent speech, and also dyspraxia, which hampers writing (Ming, Brimacombe & Wagner, 2007).



DOVE È LA CINTURA DI ASTEROIDI ?

FRA MARTE E GIOVE SÌ 2° per

FRA IL SOLE E MERCURIO NO 1° per

FRA LA TERRA E IL SOLE

FRA GIOVE E SATURNO

COME SONO LE COMETE ?

HANNO UNA LUNGA CODA FATTA DI GHIACCIO E POLVERE SÌ

SONO ROCCE CHE VAGANO NELLO SPAZIO

SONO PIANETI PIÚ PICCOLI

SONO STELLE FORMATE DA ALTRE STELLE

Prova significativa 6,3/10

BRAVO!

Fig. 2

**Figure 2.** *Test (in Italian) on the solar system, with four alternatives and up to two corrections allowed (i.e. three attempts). The student, with a teacher in attendance, went through each topic and then answered the related questions, each of which was read aloud (both the question and the alternatives) by the teacher. The student was left to mark the answer with no prompt and, if the answer was wrong, was invited to pay more attention and to try once again. On 10 questions, 6 were answered correctly at the first attempt, 2 at the second, and 1 at the third, while one was missed with no attempt (the whole page was doodled). The test was completed by a 16-year-old boy with nonverbal autism and Down syndrome. The test was statistically significant (p=0.01) and the grade was 70/100 (not 6.3/10 as on the picture).*

*[Translation: (1st page) WHERE IS THE ASTEROID BELT? A) Between Mars and Jupiter (YES, chosen at the 2nd attempt), B) Between the Sun and Mercury (NO, chosen at the 1st attempt), C) Between Earth and the Sun, D) Between Jupiter and Saturn. (2nd page) WHAT DO COMETS LOOK LIKE? A) They have a long tail, composed of ice and dust (YES), B) They are rocks wandering through the space, C) They are small planets, D) They are stars, composed of other stars. The test is significant 6.3/10. WELL DONE!]*

The need for flexible valuation tools in academic testing is evident in Italy, where since 1977 all students with any kind of disadvantage (from challenging behavior to clinical disability) of any severity (from dyslexia to severe autism) must attend, by law, mainstream schools. Where

necessary, support teachers help class teachers to provide an individualized program, and to integrate the student in class activities. In this environment, it is important to be able to evaluate the academic achievements of students with severe disabilities, who might be unable to speak or write in an independent way. Multiple-choice tests seem to be appropriate to this purpose, allowing choices to be made based on objects or pictures (figure 1), written words or sentences (figure 2).
How can we know that a result in a multiple-choice test is statistically significant, i.e. the likelihood that it was wholly obtained by choosing at random is low enough to support strong conclusions? If a student gives a wrong answer, is it possible to allow him/her to try again? If we do allow extra attempts, how does the test's significance change, and how can we mark it? The aim of the following applet is to answer these questions, offering a reliable valuation of MC test scores when extra attempts are allowed. Unfortunately, prevailing attitudes toward people with severe disabilities (such as nonverbal autism) are often rather extreme; either to believe that they understand nothing or to believe that they understand everything but are unable to show it.

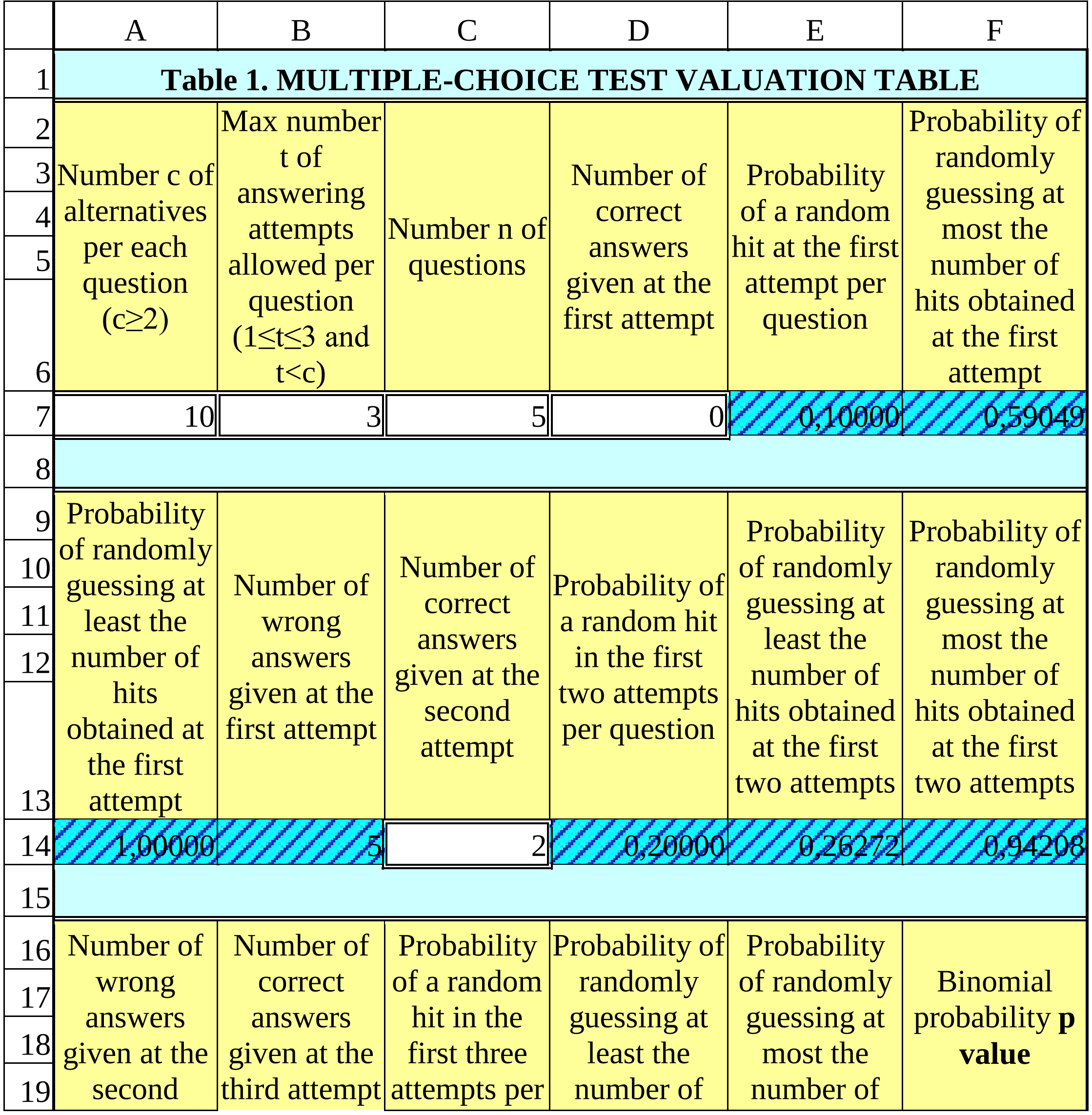
Neither of these attitudes is helpful, since particular students might confirm either or both preconceptions when tested in different ways. For this reason, the student has to be prepared on the topic on which he/she is tested, and the teacher has to be sure that each question is intelligible to him/her. In practice, it is important to write the items that have to be read aloud, providing both visual and auditory presentation for each question. For students with severe autism, the task of choosing might not be straightforward and they might need a specific training, using some physical support at the beginning, due to their impairment in learning by imitation (Rogers & Williams, 2006). In fact, choosing by pointing might be a poor method in this case, as these students often touch everything: it may be more effective to ask the student to place the chosen item inside a box or on the hand of the teacher (Schopler & Mesibov, 1995). If the choices are either written or displayed as pictures on paper, a simple tick may be a less reliable response that the coloring of a corresponding box (fig.1 and 2). Unfortunately, these students are often much more stressed than others are by the testing environment, and fail to consistently attend the test items while making their choices (Twachtman-Cullen, 2006). In some studies, visual attention has been described as a key strength (O'Riordan & Plaisted, 2001) of people with autism; that may be so, but impairments in the shifting of attention (deficits in executive functions) can undermine that strength by making it difficult to focus the student’s attention towards a new task (South, Ozonoff, McMahon, 2007; Courchesne et al., 1994). For this reason, if the answer is wrong, it can be useful to say and/or to write "no" close to the wrong answer and to immediately invite the student to rethink the question, reading it again with him/her, and inviting the student to answer it once more (see Figures 1 and 2).

With the following applet, up to three attempted answers (i.e. two corrections) are allowed per question. It may be useful to write a note (1st, 2nd, 3rd) indicating the order of the answers, as in Figures 1 and 2, to understand what criteria (if any) were used in making each choice. Note that students are allowed to mark only one alternative at a time and should be able to exclude the alternatives that they have already chosen, because they have access to immediate feedback from their previous mistakes. For this reason, not all MC tests are suitable for corrections. In fact, no correction should be allowed for stems associated with just two alternatives and stems associated with three alternatives should permit at most one correction (i.e. two attempts). At least as currently defined, our method permits at most two corrections (i.e. three attempts) for any stem associated with four or more alternatives.

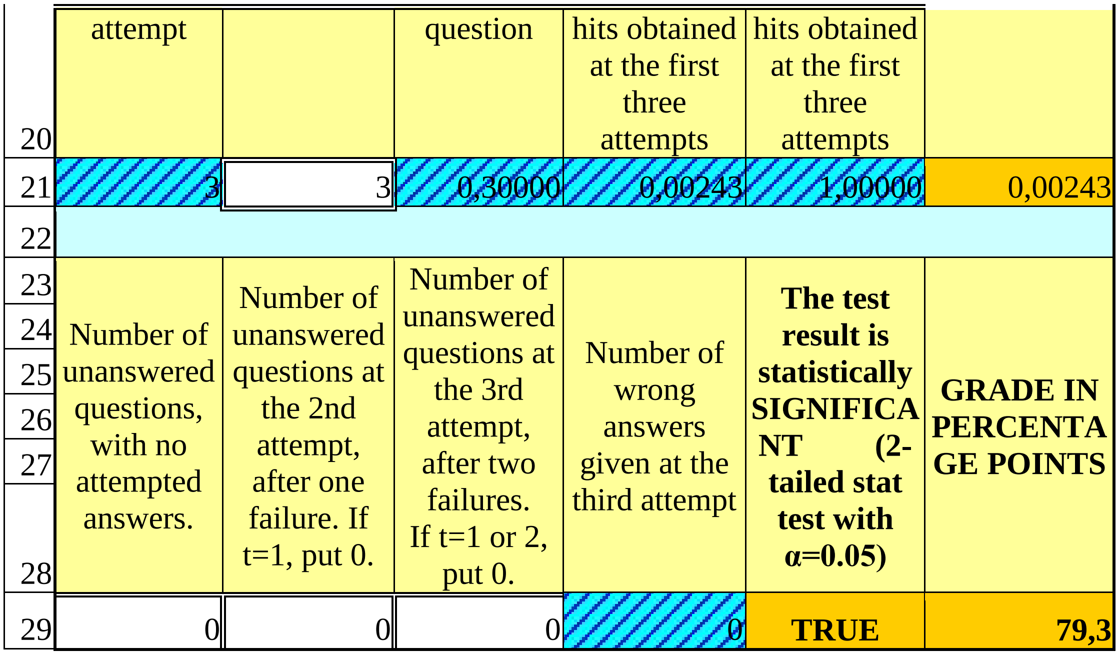
The aim of this paper is to describe a tool that also permits the reliable evaluation of students with severe disabilities, which is robust to disabilities that implicate the focusing of attention, even when the students exhibit an oppositional defiant behavior during the test. For instance, if all the answers are wrong in a two-alternative test with 6 questions, it is unlikely that the student is choosing at random (p=0.0156), but much more likely that his/her errors are really a form of protest.

Both with typical students and in special education, MC testing may be a useful tool for testing academic achievements, but should never be the only testing method employed. In fact, performance tests allow students greater freedom to produce original ideas and to supply their own information, though they may be more difficult to learn and complete. These tests are therefore at least as important as the MC tests with which we are concerned.

**The Multiple Choice Test Valuation Applet.**



| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | **Table 1. MULTIPLE-CHOICE TEST VALUATION TABLE** | | | | | |
| 2 | Number c of alternatives per each question (c≥2) | Max number t of answering attempts allowed per question (1≤t≤3 and t<c) | Number n of questions | Number of correct answers given at the first attempt | Probability of a random hit at the first attempt per question | Probability of randomly guessing at most the number of hits obtained at the first attempt |
| 3 | | | | | | |
| 4 | | | | | | |
| 5 | | | | | | |
| 6 | | | | | | |
| 7 | 10 | 3 | 5 | 0 | 0,10000 | 0,59049 |
| 8 | | | | | | |
| 9 | Probability of randomly guessing at least the number of hits obtained at the first attempt | Number of wrong answers given at the first attempt | Number of correct answers given at the second attempt | Probability of a random hit in the first two attempts per question | Probability of randomly guessing at least the number of hits obtained at the first two attempts | Probability of randomly guessing at most the number of hits obtained at the first two attempts |
| 10 | | | | | | |
| 11 | | | | | | |
| 12 | | | | | | |
| 13 | | | | | | |
| 14 | 1,00000 | 5 | 2 | 0,20000 | 0,26272 | 0,94208 |
| 15 | | | | | | |
| 16 | Number of wrong answers given at the second | Number of correct answers given at the third attempt | Probability of a random hit in the first three attempts per | Probability of randomly guessing at least the number of | Probability of randomly guessing at most the number of | Binomial probability **p value** |
| 17 | | | | | | |
| 18 | | | | | | |
| 19 | | | | | | |



| | | | | | | |
|---|---|---|---|---|---|---|
| 20 | attempt | | question | hits obtained at the first three attempts | hits obtained at the first three attempts | |
| 21 | 3 | 3 | 0,30000 | 0,00243 | 1,00000 | 0,00243 |
| 22 | | | | | | |
| 23<br>24<br>25<br>26<br>27<br>28 | Number of unanswered questions, with no attempted answers. | Number of unanswered questions at the 2nd attempt, after one failure. If t=1, put 0. | Number of unanswered questions at the 3rd attempt, after two failures. If t=1 or 2, put 0. | Number of wrong answers given at the third attempt | **The test result is statistically SIGNIFICANT (2-tailed stat test with α=0.05)** | **GRADE IN PERCENTAGE POINTS** |
| 29 | 0 | 0 | 0 | 0 | **TRUE** | **79,3** |

**Table. 1.** *Multiple-choice test valuation table of the Excel applet. To use it, replace the actual numbers of all white cells with your numbers and click, with the left key of the mouse, on the "significant" cell and on the "grade" cell. The results, in the golden cells, will appear updated with your data. If "error" appears, this means that your data are not consistent and you have made some mistake when entering them.*

The applet (Table1), programmed in Excel (see appendix), assesses the probability that the overall result of the test was achieved only by random guessing [for the probabilistic terminology, see (Spiegel, Schiller, Srinivasan, 2000)].

The logic of the calculation stems from the following claim:

> *if c is the number of alternatives and t is the (fixed) number of attempts allowed per question, with 0< t <c, then, for each stem, (i) the probability of failure in all the t attempts allowed is 1-(t/c) and (ii) the probability of guessing the correct answer in at most t attempts, is t/c.*

[Proof: The probability that the student fails in all the *t* attempts (i) is

$$\left(1-\frac{1}{c}\right)\left(1-\frac{1}{c-1}\right)\left(1-\frac{1}{c-2}\right)....\left(1-\frac{1}{c-t+2}\right)\left(1-\frac{1}{c-t+1}\right) =$$

$$=\left(\frac{c-1}{c}\right)\left(\frac{c-2}{c-1}\right)\left(\frac{c-3}{c-2}\right)....\left(\frac{c-t+1}{c-t+2}\right)\left(\frac{c-t}{c-t+1}\right)= \frac{c-t}{c}= 1-\frac{t}{c}$$

and then the probability of a correct guessing in at most t attempts (ii) is $1-\left(1-\frac{t}{c}\right)=\frac{t}{c}$. In fact, the probability of guessing the correct answer, exactly at the t-th attempt, after t-1 previous failures (conditional probability) is $\frac{1}{c-(t-1)}=\frac{1}{c-t+1}$, because we suppose that the student eliminates the t-1 wrong alternatives he has previously chosen. Hence the probability of failing even the t-th attempt, after t-1 failures, is $1-\frac{1}{c-t+1}$ for t = 1, 2,…, c-1.]

In this claim, we suppose that, at every new attempt, all alternatives are equally likely and that the student excludes all wrong alternatives (if any) that have previously been chosen. If that latter assumption is incorrect – if a student does not exclude options that have been chosen before and confirmed be wrong – his/her probability of random success will be lower, so our calculation of an upper bound on the probability of making correct choices will remain unaffected.

Hence, if *c* is the number of alternatives, then the probability of guessing exactly *h* correct answers out of *n* questions, allowing at most t attempts per question, with 0< *t* <*c*, is

$$p =\frac{n!}{h!(n-h)!}\left(\frac{t}{c}\right)^{h}\left(1-\frac{t}{c}\right)^{n-h}$$

which is the result of the binomial probability density function, BINOMDIST (non cumulative) in Excel. But the exact probability value is not useful as soon as the number of questions is high: for instance in a two choices test the exact probability of guessing 150 questions out of 300 is 0.0460, which is less than α = 0.05 (the usual limit of the statistical significance of a test), while 150 = 300/2 is the expected value of the random guessing. If we use the multinomial probability density formula[1] in a four choices test with 12 questions, the probability of getting exactly 3 correct answers at the first attempt, 3 at the second and 3 at the third, is 0.0220, while these results are exactly the expected values with the random guessing. Another problem with the above formula is that a four choices test with 10 questions, 7 answered correctly at the first attempt and one answered correctly at the third, seems to have the same probability as a similar test in which 8 questions were correctly answered at the third attempt (0.2816), while the probability of getting 7 out of 10 questions at the first attempt is 0.0031. For these reasons, we consider as null hypothesis the claim 'all the hits were achieved by random guessing' and we compute the probability p of erroneously rejecting the null hypothesis (*p-value*). This *p*-value is computed as the smallest of the 2*t* values $p_k$ *and* $p'_k$, with *k* = 1, 2, …, *t* , calculated as below:

$$p_k = \sum_{i=0}^{h_k} \frac{n!}{i!(n-i)!}\left(\frac{k}{c}\right)^i \left(1-\frac{k}{c}\right)^{n-i}$$

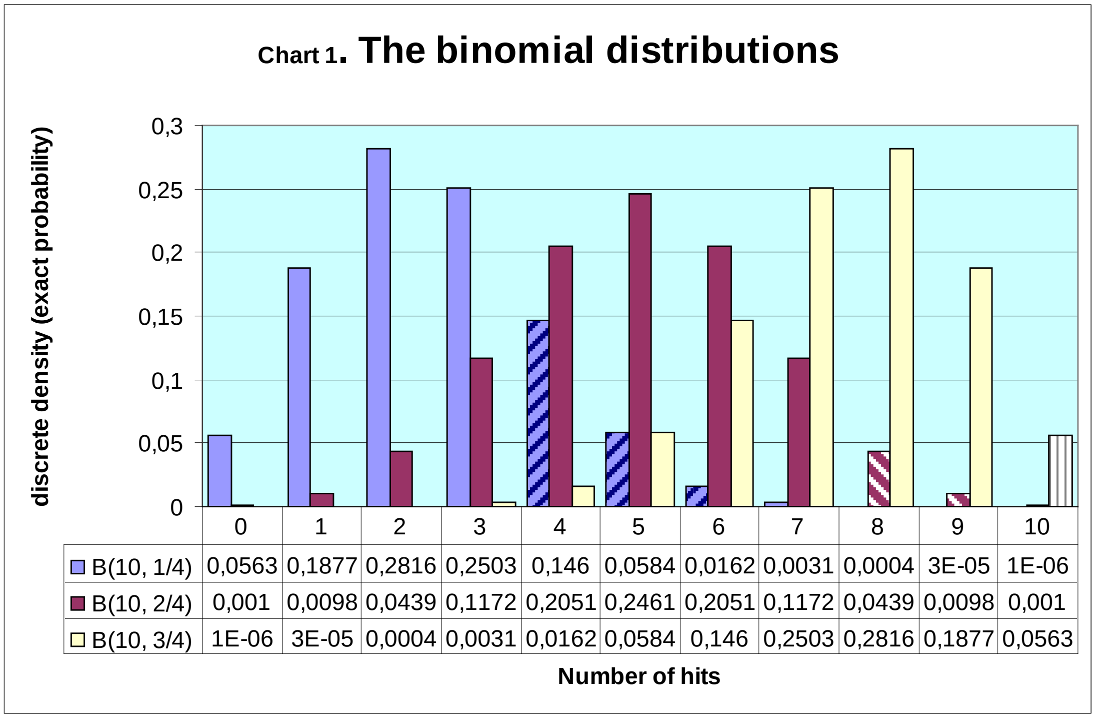
and

$$p'_k = 1 - \sum_{i=0}^{h_k - 1} \frac{n!}{i!(n-i)!}\left(\frac{k}{c}\right)^i \left(1-\frac{k}{c}\right)^{n-i}$$

with $k = 1, 2, \ldots, t$ , where $h_k$ is the number of hits obtained in at most *k* attempts and if $h_k$=0, we suppose $p'_k$,=1. For each allowed attempt *k*, these probabilities are, respectively, the probability of having at most $h_k$ hits in at most k attempts and the probability of having at least $h_k$ hits in at most k attempts. The reason for considering both $p_k$ and $p'_k$, (rather than just the latter) is that these students have been known to express their opposition to a test (or to a teacher) by making deliberate mistakes; that situation can be inferred if we observe performance scores that are significantly below what might be expected from random guessing. Hence, as it is a two tailed test, the significance limit will be $\alpha/2 = 0.025$, if $\alpha = 0.05$ is chosen, as in the applet, and if the number *n* of questions is greater than 1. If $n = 1$ the significance limit will be α.

In the applet, we suppose $t \leq 3$, because making a request for further corrections could frustrate students with difficulties in focusing. In the computation of the probability, we consider as completely failed the questions, which are unanswered either at the second or at the third attempt after one or two failures. To assess the statistical significance, in the applet, we consider the three binomial distributions *B(n,1/c)*, *B(n,2/c)* and *B(n,3/c)*, respectively of the number of hits given in the first attempt, the number of its hits given in the first two attempts and the number of its given in the first three attempts, and we look for a given result that is too far from the mean to be considered a chance result (Chart 1)

Alongside the number of attempts allowed, the number of questions and the number of choices per question also influence the significance of the test. For instance, the number of questions has to be at least three in a 4-choice test to achieve a significant result when all answers are hits at the first attempt, while in a 2-choice test, the number of questions required is at least six.



| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| □ B(10, 1/4) | 0,0563 | 0,1877 | 0,2816 | 0,2503 | 0,146 | 0,0584 | 0,0162 | 0,0031 | 0,0004 | 3E-05 | 1E-06 |
| ■ B(10, 2/4) | 0,001 | 0,0098 | 0,0439 | 0,1172 | 0,2051 | 0,2461 | 0,2051 | 0,1172 | 0,0439 | 0,0098 | 0,001 |
| □ B(10, 3/4) | 1E-06 | 3E-05 | 0,0004 | 0,0031 | 0,0162 | 0,0584 | 0,146 | 0,2503 | 0,2816 | 0,1877 | 0,0563 |

**Chart 1**. *Suppose that a MC test with 4 choices, up to 3 answering attempts allowed per question, and 10 questions, yielded 4 correct answers at the first attempt, 4 at the second and 2 at the third. The binomial distribution of hits obtained at the first attempt is B(10, ¼), the binomial distribution of hits obtained in at most two attempts is B(10, ½) and the binomial distribution of hits obtained in at most three attempts is B(10, ¾). In B(10, ¼), the probability $p'_1$ of having at least 4 hits at the first attempt is $p'_1$ = 0.146+0.0584+0.0162+0.0031+0.0004+0.00003+0.000001=0.22412, which the sum of the values of the striped columns in that distribution. In B(10, ½), the probability $p'_2$ of having at least 8 hits in at most two attempts, is $p'_2$ = 0.0439 + 0.0098 + 0,001 = 0.0547 (sum of the striped columns in that distribution). In B(10, ¾), the probability $p'_3$ of having 10 hits in at most three attempts, is $p'_3$ = 0.0563 (striped column). As in each distribution, the probabilities $p_1$, $p_2$, $p_3$ of having respectively at most 4, 8 and 10 hits are much greater than $p'_1$, $p'_2$, $p'_3$ , the p value is $p'_2$ = 0.0547, which is not less than 0.025. Hence the test result is not statistically significant.*

### *How are the Grades Computed?*

In the final score, measured in percentage points, errors are computed negatively, and that negative score is added to a positive score if there is a later correction. A test can be split and answered at different times, for instance on different days, if the student is tired.

Let *c* be the fixed number of answering alternatives per question (with $c>1$) and *t* be the maximum number of answering attempts, allowed per question, with $1 \leq t \leq 3$. Choosing at random at the first attempt, the probability of a hit, in each question, is 1/*c*, and of an error is 1-(1/*c*) (Spiegel et al., 2000). Hence we decide, following (Culwick, 2002), that, at the first attempt, -1/*c* is the score for each error, and 1-(1/*c*) is the score for each hit (this removes the advantage 1/*c* given by guessing), while zero is the score for unanswered questions. At the second attempt, with

$c>2$, the score for each error is $–(1/c) –(1/(c-1))$, the score for each hit it is $1-(1/c)-(1/(c-1))$ and the score for each unanswered question is $(-1/c)$, adding a penalty $(-1/c)$, which refers to the error that was previously made, and changing the guessing penalty to $(-1/(c-1))$, since there are now just $c$-1 alternatives. At the third attempt, with $c>3$, the scores are $-1/c-(1/(c-1))-(1/(c-2))$ for each error, $1-(1/c)-(1/(c-1))-(1/(c-2))$ for each hit, and $-1/c-(1/(c-1))$ for each unanswered question, because there are now two previous errors and the number of the alternatives is reduced to $c$-2. Finally, following (Culwick, 2002), we divide the average score of the questions by $1-(1/c)$, which is the highest score, and multiply it by 100, defining 100 as the maximum score that can be achieved for each hit at the first attempt. The formula $100s/(1-(1/c))$ converts each question score $s$, computed as above, into a percentage grade (Table2). Hence, the grade for each error at the first attempt is calculated as $–100(1/c)/(1-(1/c))$, which is equal to $100(1/(c-1)$ that is the formula employed by the S2 method cited previously, when the maximum score is 100 instead of 1.

The rationale of this scoring procedure is that, in the same test, hits obtained in choosing among a certain number of alternatives are given more weight than hits obtained from choices among fewer alternatives, but conversely errors obtained in choosing among fewer alternatives are given more negative weight than errors obtained in choosing among more alternatives; in fact, guessing probabilities depend on the number of alternatives available.

Finally, the result of a test is considered credible if it yields significant divergence from a "guessing distribution", i.e. $p < 0.025$ if $n>1$ and $p<0.05$ if $n=1$, and if the student achieves a global score of at least 60%. Other criteria can also be used to define a "passing" score.

| **Number of alternatives** | **Table 2. SCORES FOR EACH ANSWER IN PERCENTAGE POINTS** | | | | | | |
|---|---|---|---|---|---|---|---|
| | Scores for each **hit** at the **1**rs attempt | Scores for each **hit** at the **2**nd attempt | Scores for each **hit** at the **3**rd attempt | Scores for each **fail** at the **1**st attempt | Scores for each **fail** at the **2**nd attempt | Scores for each **fail** at the **3**rd attempt | Scores for **not replied** questions |
| **2** | 100 | - | - | -100 | - | - | 0 |
| **3** | 100 | 25 | - | -50 | -125 | - | 0 |
| **4** | 100 | 55,56 | -11,11 | -33,33 | -77,78 | -144,44 | 0 |
| **5** | 100 | 68,75 | 27,08 | -25,00 | -56,25 | -97,92 | 0 |
| **6** | 100 | 76,00 | 46,00 | -20,00 | -44,00 | -74,00 | 0 |
| **10** | 100 | 87,65 | 73,77 | -11,11 | -23,46 | -37,35 | 0 |
| **26** | 100 | 95,84 | 91,51 | -4,00 | -8,16 | -12,49 | 0 |

**Table. 2.** *The grade of the whole work is obtained by dividing the sum of the scores for each question by the number of questions (mean score). The 10 alternatives might be the Arabic digits in a math task and the 26 alternatives might be the alphabet letters in a spelling task.*

### *Writing Tasks by Stamps: Spelling and Computing Mathematical Operations*

The applet can also be used to evaluate writing tasks (figure 4 and 5), if they are done by choosing the letters or digits in a fixed pool. To simplify the choosing task, we suggest an arrangement in which the positions of the letter and digit stamps are fixed, as in figure 3.
In general, the display of the alternatives in all tasks that require a selection among a fixed set (such as true-false tasks, colors tasks and so on) should be kept constant to avoid adding unneeded complexity to the task. The writing task can be organized incrementally, beginning with only a few letters or digits on a keyboard and adding others in further sessions. Instead of stamps, letter cards can be used, or a computer keyboard, hiding the keys that currently play no part.

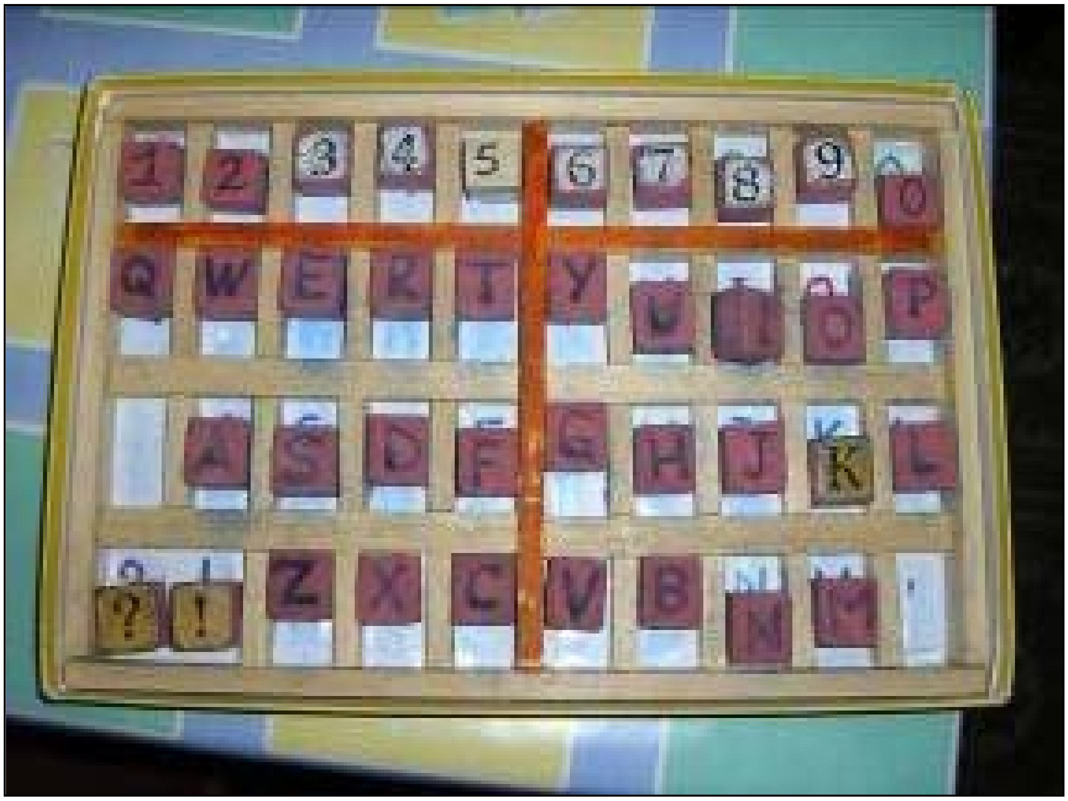


**Figure 3.** *The "keyboard", with rubber (worn-out) stamps displaying letters of the alphabet and Arabic digits, has a wooden grid that can be taken off to allow for changing and cleaning of the background; this background is printed with letters and numbers and is covered with a transparent slide. Letters and numbers are displayed as for the English computer keyboard (QWERTY), to aid the student in selecting. In math tasks, the letters can be removed and the digits can be removed in spelling tasks.*

As we can see in Figures 4 and 5, the selection of the letters was corrected by crossing out the wrong letter and writing a small "1" or "2" or "3" close by (or by using three different color ink pens), to indicate if it was the first, the second or the third error. After the third error, a third correction is not allowed and the selection is counted as an error at the third attempt. These tasks are often statistically significant, because there are many alternatives (26 and 10) for each selection, and they can also help us to understand what these students really know, going beyond any initial prejudices.

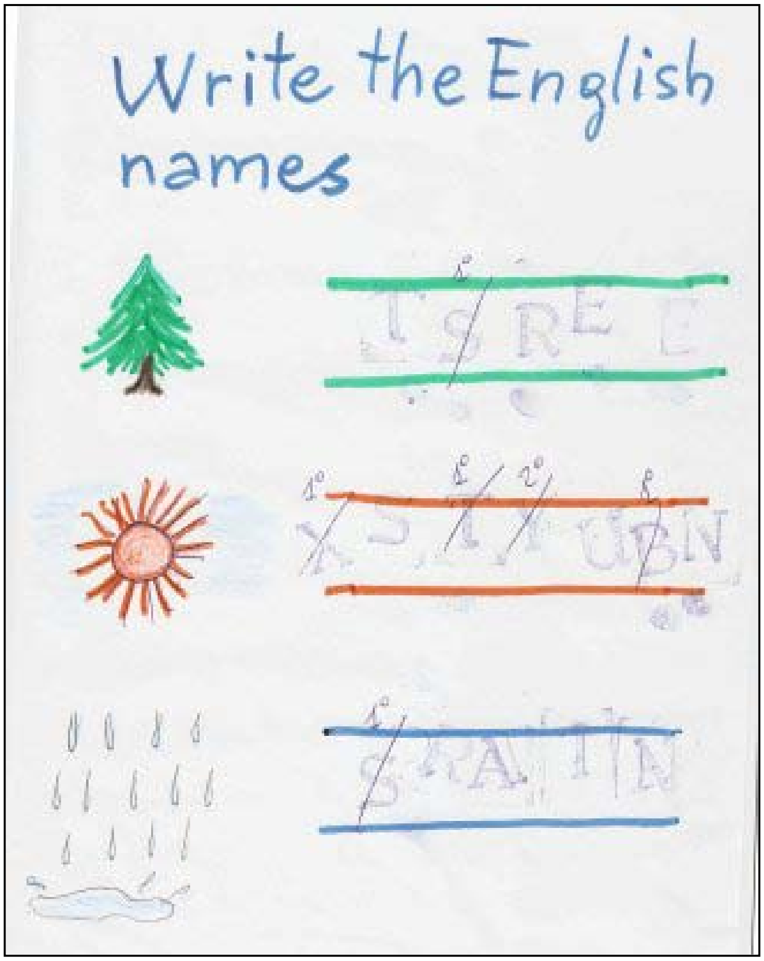
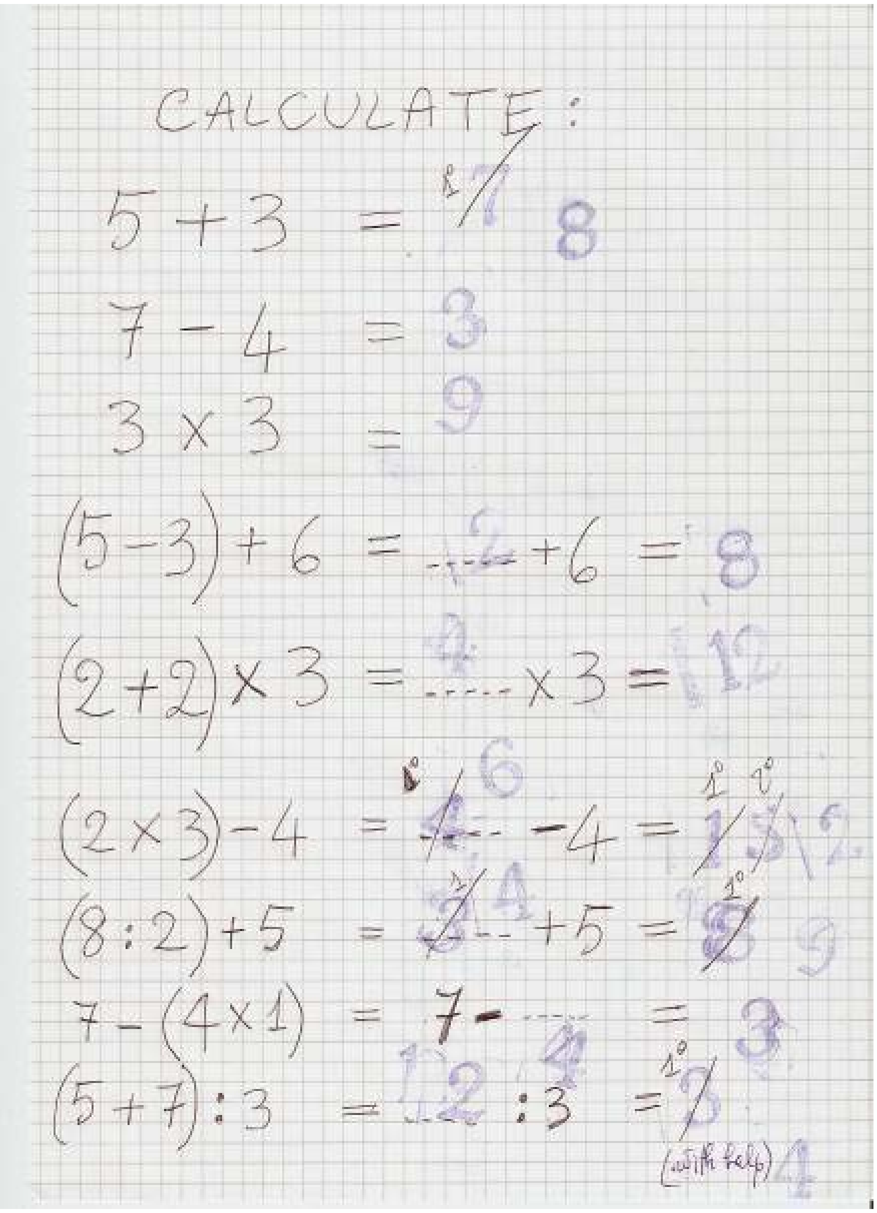


**Figure 4**. *This is an example of a spelling task, where up to two corrections are allowed per letter selection. When the selection was wrong, the student was informed and the wrong letter was crossed out, indicating if it was the first, the second, or the third error for that selection (three different colored pencils could be used to distinguish the order of the errors). The corrected new letter was printed on the left. In this exercise, the words TREE, SUN and RAIN should be written. The student, a 17-year old boy with Down syndrome and autism, educated in Italian mainstream schools, was not helped in the selection, but was helped in stamping – in deference to his severe dyspraxia. Out of 11 letters, he selected 6 of them correctly at the first attempt (T, E, E, A, I, N), 4 at the second attempt (R, S, N, R) and 1 at the third attempt (U). Hence, using the applet, we find that the test was significant (p<0.0001) and the grade was 97.7%. This grade is high, because the penalty for the errors is low, i.e.(1/26)=0.038. Observe that the mistaken letter stamps were close to the correct ones on the keyboard, and the selection of "U" was achieved in three steps.*



CALCULATE:

5 + 3 = ~~7~~ 8

7 − 4 = 3

3 x 3 = 9

(5 − 3) + 6 = 2 + 6 = 8

(2 + 2) x 3 = 4 x 3 = 12

(2 x 3) − 4 = ~~4~~ 6 − 4 = ~~1~~ ~~3~~ 2

(8 : 2) + 5 = ~~3~~ 4 + 5 = ~~8~~ 9

7 − (4 x 1) = 7 − 4 = 3

(5 + 7) : 3 = 12 : 3 = ~~3~~ 4

(with help)

**Figure 5.** *This math task was carried out by the same student and up to two corrections are allowed per selection, as in the previous example. Out of 17 digits that should be printed, 11 were selected correctly at the first attempt, 4 were selected correctly after one correction, 1 was selected correctly after two corrections, and 1 was wrong (in fact, it was selected with help - so is not counted). In this case the number of alternatives is 10 and the grade was 87.5% - a significant result (p<0.0001). This test could also certainly be repeated with other grading criteria.*

For instance, in the example of figure 4, if task were completed without allowing corrections, the result would be “TSEE” for TREE, “*XTB*” for SUN and “*S*AIN” for RAIN, which is not very promising. In the same condition, the example of figure 5 would be 5+3=*7*, 7−4=3, 3×3=9, (5−3)+6=2+6=8, (2+2)×3=4×3=12, (2×3)−4=*4*−4=*1*, (8:2)+5=*3*+5=8, 7−(4×1)=7−4=3, (5+7):3=12:3=*3*. In the spelling task, no word was correctly written at the first attempt, and the percentage of correct letters was 54.5%. The math task was better and the percentage of correct digits was 64.7% at the first attempt. With our grading criterion, which penalizes the errors, the grades should be 52.7% and 60.8%, respectively, if no correction was allowed. If we observe the errors, we see that almost all the mistaken letters and digits were close to the correct ones on the keyboard. Allowing corrections helps the student to be more confident in his/her own ability - despite his/her impairments - and to rethink the task, as well as helping the teacher to better understand the cause of the errors. Using the grading criteria described above for multiple-choice tests, we employed the applet to compute the grades in the two tasks, with two corrections allowed, as 97.7% and 87.5% respectively. Our grading choice does not exclude further grading criteria that are more frequently used for these tasks. The high statistical significance, with $p<0.0001$ in both tasks, confirms that the choices were not usually guessed. The capacities and difficulties of a particular student cannot be generalized to other students with the same pathology; each student might have gone through different learning experiences, and in any case, apparently similar disorders can emerge from damages to very different brain areas. Recognizing this, we suggest that task structures should be flexible, and adapted individually to each student.

## *About Grading and Significance*

As we have seen, negative grades are a useful way to counterbalance any benefits that accrue from random guessing (Culwick, 2002); they should, however, not be exposed to students directly, who should simply be told that they have not passed the test.

Even a test that is not statistically significant can yield useful information, because it may be significant from other points of view. For instance, if a student with autism’s behavior improves while dealing with a new topic (e.g. displaying a reduced tendency toward head banging), we can conclude that the student might be interested in that topic. If, moreover, he accepts to participate in a test, this is truly a positive signal of pleasure and then it does not matter if his test ‘p value’ exceeds 0.025, because the statistical significance will improve later, when the topic will be more familiar.

Here we want to emphasize the role of chance in a multiple-choice test, and how arbitrary some judgments (either positive or negative), which are made after only a few observations with few alternatives per question, may be. A ‘quick check’, with either one or two questions, and with two or three alternatives per question, tell us virtually nothing (either positive or negative) about students’ achievements. In fact, even correct responses may tell us nothing reliably positive, because, for instance in the case of a two-alternative test, at least six correct answers must be observed, with no previous error, before we can believe that the student has learned and is really making deliberate choices. On the other hand, consistently incorrect answers may tell us nothing reliably negative, because students with attention deficit and behavioral problems are likely, even in an informal testing situation, to reply initially by choosing at random; for this reason, here we propose to give them the chance to make corrections. The policy suggested is to be more patient and respectful with these students, giving them more opportunities to show their knowledge.

We can observe that even a single positive answer, without previous errors, is statistically significant if the alternatives are 26, as when students select from among the letters of the English alphabet. For example, if the question is “What is the initial letter of the capital of the U.K.?”, and the answer is “L”, the probability of guessing with 26 letters is 0.038, which is lower than 0.05 and therefore significant, according to our criteria. On the other hand, in the same kind of task with 26 alternatives, the student can fail the first 4 letters at the first attempt, but the test is still statistically significant if there are two further correct letters at the first attempt (p=0.020). For instance, if the task is “Write on the computer the names of the animals in the pictures” and the student, instead of writing CAT and DOG, writes *PEN* and *R*OG, without corrections, then the test is statistically significant (p=0.020 ), even if the grade is too low to pass (30.7%). This does not mean that the student did not guess, because much more improbable events may happen, such as winning the lottery or having Down syndrome and autism (less than 1 out of 6000, i.e. p=0.000167). Since we cannot know for certain if the student is simply guessing, judgments of this sort are best made by following a fixed criterion.

### The Use of the Applet with a Nonverbal Student with Autism and Down Syndrome

The applet was used to evaluate some of the knowledge of a 16-year-old nonverbal student with severe autism spectrum disorder and Down syndrome, attending a secondary school in Italy. He seemed to understand and to remember some topics in science, history, history of arts, Italian literature and English. In particular, he liked to read (with the help of his support teacher) the same novels that his classmates were set, and seemed to understand them well, as tested through his responses to multiple-choice tests allowing up to two corrections per question. These tests usually employed four alternatives per question (Figures 1 and 2). Applying the method reported here, the teachers were able to grade his progress in most of the courses that he attended.

## Conclusion

The Excel applet and the method of allowing some possibilities of correction in multiple-choice tests, which are proposed here, need to be verified in further practical trials before we can be sure of their real utility. Regardless of the results of that work, the applet might play a more general role in helping testers to assess the role of guessing in students' scores for multiple-choice tests (with or without corrections) and true/false tests. The applet also computes percentage grades, employing a formula that penalizes errors (if *c* is the number of alternatives, the penalty is -100/(*c*-1), for questions replied at the first attempt) and balances the value of any subsequent hits and errors against the number of residual alternatives. An extension of the use of the applet is proposed both for spelling and mathematics tests (with or without corrections) that are carried out printing the words and the numbers with stamps displayed on a keyboard or with the computer or composing them with letter/digit cards selected from a fixed set. Employing this method, we were able to test a 16-year-old nonverbal student, with autism spectrum disorder and Down syndrome, effectively, identifying his strengths and weaknesses, as well as some of his preferences, and discovering that he was both able and eager to share the learning culture of his peers in a mainstream secondary school, in spite of his limits.

## Acknowledgements

In the Dept. of Pure and Applied Mathematics, University of Padua (Italy) I wish to thank prof. Paolo Dai Pra for his useful remarks and prof. Paolo Malesani, for our talk on probability. I wish also to thank the teachers of the secondary school “Istituto d’Arte P. Selvatico” of Padua and Pamela Spada for using multiple-choice tests and the applet with a student with autism and Down syndrome. I thank also Thomas Hope for his help in improving the English language style of this paper.

## *References*

Burton, S.J., Sudweeks, R.R., Merrill, P.F., Wood, B. (1991*) How to prepare better Multiple-Choice Test Items: Guidelines for University Faculty*. Provo, UT: Brigham Young University Testing Services.

Courchesne E, Townsend J, Akshoomoff NA, Saitoh O, Yeung-Courchesne R, Lincoln AJ, James HE, Haas RH, Schreibman L, Lau L. (1994). Impairment in shifting attention in autistic and cerebellar patients. *Behavioral Neuroscience*, 108 (5), 848-865.

Culwick, M. (2002). *Designing & Managing MCQ's*. Handbook. http://www.le.ac.uk/castle/resources/mcqman/mcqcont.html retrieved on 30 Nov. 2007.

Haladyna, T. M., Downing, S.M., Rodriguez, M.C., (2002). A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment, *Applied Measurement in Education*, 15(3), 309–334.

Heuer, S., & Hallowell, B. (2007). An evaluation of multiple-choice test images for comprehension assessment in aphasia. *Aphasiology*, 21 (9), 883 – 900.

Martinez, M. (1999). Cognition and the question of test item format. *Educational Psychologist*, 34, 207–218.

Ming, X., Brimacombe, M., & Wagner, G.C. (2007). Prevalence of motor impairment in autism spectrum disorders. *Brain and Development*, 29 (9), 565-570.

O'Riordan, M.A.F. & Plaisted, KC (2001). Enhanced discrimination in autism. *Quarterly Journal of Experimental Psychology*, Section A, 54, 961-979.

PPVT™-4: *Peabody Picture Vocabulary Test, Fourth Edition*, (2006) Pearson Education, Inc.

Prieto, G., & Delgado, A.R., (1999). The effect of instructions on multiple-choice test scores. *European Journal of Psychological Assessment*, 15 (2), 143–150.

Rogers, S. J., Williams, J.H. (2006) Imitation in Autism: Findings and Controversies. In S. J. Rogers & J. H. Williams (Eds), *Imitation and the Social Mind: Autism and Typical Development*, Guilford Press, New York.

Schopler E. & Mesibov G.B. (Eds.) (1995), *Learning and Cognition in Autism*, New York: Plenum Press.

Spiegel, M.R., Schiller, J., Srinivasan, R.A. (2000) *Schaum's Outlines of Theory and Problems of Probability and Statistics,* 2nd Edition, McGraw-Hill Companies Inc., USA.

South, M., Ozonoff, S., McMahon, W.M. (2007). The relationship between executive functioning, central coherence, and repetitive behaviors in the high-functioning autism spectrum. *Autism*, 11 (5), 437-451.

Thurstone, L.L. (1919). A method for scoring tests. *Psychological Bulletin*, 16, 235–240.

Twachtman-Cullen, D. (2006). Communication and stress in students with autism spectrum disorders. In Baron, M. G., Groden, J., Groden, G., Lipsitt, L. (Eds). *Stress and coping in autism*. (pp. 302-323). New York, NY, US: Oxford University Press.

Volkmar, F.R., Paul, R., Klin, A.., Cohen, D. (Eds) (2005). *Handbook Of Autism And Pervasive Development Disorders*, (3rd edition) John Wiley & Sons Inc.

## *Appendix*

The applet in table1 was programmed in Excel with the following formulas:

```
E7 =1/A7
F7 =BINOMDIST(D7;C7-A29;E7;TRUE)
A14 =IF(D7>0;1-BINOMDIST(D7-1;C7-A29;E7;TRUE);1)
B14 =C7-A29-D7
D14 =IF(B7>1;2/A7;E7)
E14 =IF(D7+C14>0;1-BINOMDIST(C14+D7-1;C7-A29;D14;TRUE);1)
F14 =BINOMDIST(D7+C14;C7-A29;D14;TRUE)
A21 =IF(B7>1;B14-C14-B29;0)
C21 =IF(B7>2;3/A7;D14)
D21 =IF(D7+C14+B21>0;1-BINOMDIST(D7+C14+B21-1;C7-A29;C21;TRUE);1)
E21 =BINOMDIST(D7+C14+B21;C7-A29;C21;TRUE)
F21 =IF(AND(B7>0;B7<=(A7-1);B7<4;D7<=(C7-A29);C14<=B14-B29;B21<=A21-
        C29);MIN(F7;F14;E21;A14;E14;D21);"ERROR")
D29 =IF(B7=3;A21-C29-B21;0)
E29 =IF(C7>1;IF(F21<0,025;TRUE;FALSE);IF(F21<0,05;TRUE;FALSE))
F29=IF(OR(AND(B7=1;A7>1;C14=0;B21=0;B29=0;C29=0;C7>=(D7+A29));AND(B7=2;A7>2;
        B21=0;C29=0;C7>=(D7+C14+A29+B29));AND(B7=3;A7>3;C7>=(D7+C14+B21+A29+
        B29+C29)));(((1-(1/A7))*D7)+((1-(1/A7)-(1/(A7-1)))*C14)+(IF(A7>2;1-(1/A7)-(1/(A7-
        1))-(1/(A7-2));0)*B21)+IF(B7=1;(-(1/A7))*B14;IF(B7=2;(-(1/A7)-(1/(A7-
        1)))*A21;IF(A7>2;(-1/A7-(1/(A7-1))-(1/(A7-2)))*D29;0)))+(A29*0)+(B29*(-
        1/A7))+(C29*(-(1/A7)-(1/(A7-1)))))*100/((C7)*(1-(1/A7)));"ERROR")
```