

The probability distribution of the response times in a self-paced continuous search task

*Ad van der Ven*¹, *Risto Hotulainen*² & *Helena Thuneberg*³

Abstract

When psychologists began to use intelligence tests, they also used simple, overlearned tasks to determine the pattern of individual reaction times (RT). Measures of RT variation were proposed as possible indicators of intelligence. However, a fundamental question has remained partly unanswered: Is there an existing theory that explains individual RT variation? In this paper, a theory is proposed for the response times obtained in the Attention Concentration Test. The test consists of two different conditions: a fixed condition and a random condition. For each of these two conditions a different RT model was developed both based on the basic assumption that the individual response times have an approximately shifted exponential distribution. Empirical data was obtained from two different samples ($N = 362$, $N = 334$) of Finnish students. The method used to check the validity of each model involved computing the intercept and slope of the linear regression of the standard deviation from the stationary response times on the mean corrected for shift. In this regression analysis, the standard deviation is the dependent variable and the mean corrected for shift the independent variable. The shift parameter was estimated by using the smallest reaction time. The observed intercept and slope were compared with the predicted intercept and slope according to the proposed models. The model for the fixed condition of the test did not hold. The model for the random condition, however, did. The findings were interpreted according to the arrangement of the targets as they occurred in each bar.

Keywords: Exponential distribution, Erlang distribution, Solleveld distribution, Solleveld-Erlang distribution, continuous performance tests, attention concentration test

¹ Correspondence concerning this article should be addressed to: Ad van der Ven, PhD, Learning and Development, Department of Pedagogy, Radboud University Nijmegen, Montessorilaan 3, 6500 HE Nijmegen, the Netherlands; email: a.vanderven@pwo.ru.nl

² Centre for Educational Assessment, University of Helsinki, Finland

³ Centre for Educational Assessment, University of Helsinki, Finland

Introduction

At the beginning of the twentieth century, the first intelligence test was proposed by Binet. The test is known as Binet's intelligence scale. In addition to intelligence tests, however, eminent psychologists, such as Binet, Hylan and Spearman started using so-called prolonged work or continuous performance tasks (CPTs), in which subjects are required to engage in simple, repetitive activities, such as letter cancellation, detecting differences in simple shapes, adding three digits and so on. The obtained series of response (or reaction) times made it possible to examine the patterns of reaction times (Binet, 1900; Hylan, 1989). In this way, it was possible to study "the fluctuations which always occur in any persons continuous output of mental work, even when this is so devised as to remain of approximately constant difficulty" (Spearman, 1927, p. 320).

Many years later, in the Netherlands, the Bourdon-Wiersma test (see Huiskamp & de Mare, 1947) was introduced, and subsequently a test version was designed for children. This test is known as the Bourdon-Vos Test (Vos, 1992). Such tests provide several measures that can be used as indicators of the mental ability of the testee. The simplest measure is the amount of time used (speed) to accomplish the test. Furthermore, because of the nature of such tests, multiple RTs are obtained for each subject. These intra-individual RTs are used to calculate the mean or median and the standard deviation in RT over series. For example, in earlier studies on Bourdon-like cancellation tests, the preferred measure of performance fluctuations was the percentile range, defined as the difference between fastest and slowest row(s) of the test. However, the ability to use different measures has caused some debate among researchers. For example, over a century ago Hylan (1898) and Binet (1900) stressed the importance of the fluctuation in the reaction times and suggested the mean deviation as a measure of performance.

Several decades later, Spearman considered that oscillation was a separate universal factor in addition to what he called the general factor and perseveration (Spearman, 1927, p. 327). Indeed, Larson and Alderton (1990) reported the following: "Jensen (1982), discussing his reaction time (RT) experiments, noted that trial-to-trial variability (the standard deviation of each subject's reaction times) frequently surpassed response speed as a predictor of intelligence. That is, low aptitude individuals were excessively variable from one RT trial to the next, relative to brighter subjects. Currently, numerous studies suggest that his observation was correct and that variability has a robust statistical relationship to intelligence." This study is in line with the abovementioned suggestion and introduces theoretically based models explaining how RTs fluctuate during simple CPTs. More recently, a validation study with a new CPT test known as the Attention Concentration Test (ACT) was reported by Hotulainen et al. (2014). This study examined

how attention (RT variation) correlated with and contributed to scientific reasoning (a modified version of Science Reasoning Tasks) and school achievement (GPA).

Although measures of variation, such as the mean deviation and the standard deviation may be intuitively appealing for RT measures, these measures lack theoretical foundations to explain observed RT fluctuations. For example, one might rightly ask what exactly is captured in the different measures which are used. This question can be only answered by an explanatory theory of the fluctuations in the RTs. Hence, the aim of this study first is to introduce a theoretically sound conceptualization for the random variability of the RTs in the stationary part of the time series and second, to determine by testing whether the models chosen to explain the random variability of the RTs holds with the empirical data.

Attention and attention concentration tests

In this study, attention is understood as a fundamental attentional capacity (Smit & van der Ven, 1995). This definition is based on several assumptions, such as attention is present in all cognitive tasks requiring mental effort, such as CPT occurrences, and tasks requiring generic abilities, such as reasoning (Hotulainen et al., 2014). In the early 19th century, psychologists who were studying mental ability (e.g., Godefroy, 1915; Spearman, 1927) noticed that even in easy repetitive tasks (e.g., adding digit numbers, letter cancellation or crossing out instructed groups of dots) that were timed series wise, attention varied not only between subjects but also during the subjects' performance (within series). To theorize this phenomenon Pieters and Van der Ven (1982) introduced the explanatory concept of distraction to account for the RT fluctuations occurring in CPT's. They assumed that the manifest RTs over items or a series of items should be considered composed of a relatively constant real total work time, interrupted by a series of random distraction times. These assumptions led to the refinement of the inhibition theory (IT) (van der Ven, Smit & Jansen, 1989; Smit & van der Ven, 1995). The IT is a useful theoretical model to explain how RT fluctuates in prolonged overlearned response tasks, and how this is related to cognitive performances, such as intelligence (van der Ven, Gremmen & Smit, 2005). The IT is based on the basic assumption that during the performance of any mental task requiring a minimum of mental effort, the subject actually goes through a series of alternating latent states of distraction (non-work 0) and attention (work 1), which cannot be observed and are completely imperceptible by the subject (Smit & van der Ven, 1995). Thus, in a test situation, it is hypothesized that when the subject takes control of his or her progress through given test items in easy prolonged over-learned response-based tasks, the acquired measures, especially RT standard deviation, capture this and accordingly provide an indicator of the subject's

mental ability (Larson, & Alderton, 1990; Klauer, 1993; Smit & van der Ven, 1995; Steinborn, Flehmig, Westhoff, & Langer, 2008).

With the introduced theoretical preconditions of Huiskamp and de Mare (1947) and Vos (1992) in mind, the first author of this article developed the computer-based Attention Concentration Test (ACT). The ACT consists of various easy tasks, such as colour detection, counting the number of dots on an ordinary dice and determining the position of the dots of a dice with only two dots. The default version of the test consists of a series of 25 bars of buttons, some of which have to be clicked on. This results in a series of 25 reaction times, which form the basis of the final score. The ACT deviates from existing sustained attention tests, such as the Bourdon-Wiersma test and Bourdon-Vos test, in two ways: First, no errors are allowed. It is a well-known fact that subjects are able to exchange speed for accuracy. Routine tasks could be done faster, but at the cost of accuracy. The subjects are able to work accurately but at the cost of speed. This phenomenon is referred to as the speed-accuracy trade-off and can be described in terms of the so-called speed-accuracy trade-off function, in which the probability of a correct answer is given as a function of the time needed to answer the item. It is a monotone increasing function, and it may vary from subject to subject. Depending on the subject, the function's graph may be shifted to the right (a lower ability) or to the left (a higher ability). As long as one does not know the ability of the subject, one cannot know whether the observed response times are higher or lower when they are corrected for errors. This is only possible when the trade-off function of the subject in question is known, which, however, is generally not the case. In the ACT, the entire problem has been circumvented by only accepting test results without any error. Second, the test can be administered repeatedly because of the computerized bar generation. This means that the positions of the buttons are randomly displaced in each bar. In this way, the learning factor can be eliminated by giving the testee ample time to practice the test before actually taking it. In all existing intelligence tests, practice during the test may play a very important role in the final test score. Some testees may get used to the task quickly while others get used to the task more slowly. It may well be the case that although some subjects may differ in their rate of learning, they may be the same in their performance when the progress in learning has ended. Thus, if practice were allowed before actually taking the test, those subjects would have obtained the same test results. If the purpose of the test is to measure the rate of learning, then the test should be constructed accordingly.

Furthermore, the Bourdon Vos test has the same number of targets in each line. However, in the Bourdon Wiersma test, the number of targets is random (across lines). The same distinction applies to the ACT. The number of targets (buttons clicked) in each bar can be taken as fixed (Stimulus-number: fixed) or as random (Stimulus number: random or random, restricted). The difference between random, restricted and random, is that

bars with no targets do not occur under random, restricted but may occur under random. The distinction between fixed and random is very important because the probability distributions of the respective response times may be different.

Distributions explaining random variability of RTs in CPT

The ACT results in a series of response times, which on the subject level usually show an upward trend and become stationary in the long run. Occasionally, a downward trend also can be observed. This may be caused by some still occurring learning effect. The test, however, should be practiced before actually taking the test. Even if the test is fully over learned increasing or decreasing response time curves may still occur. It is not possible to design the test in such a way that increasing or decreasing curves do not occur. However, for the evaluation of the test for purposes of selection and classification the final score is derived from the response times in the stationary part of the curve. How this is done is shown in the sections on reliability. The probability distributions that will be used in this paper concern only the stationary part of the response time curve. The Erlang distribution will be proposed for the fixed condition of the test and a mixture of the Erlang distribution and the Solleveld distribution for the random condition of the test.

For each condition of the test, fixed as well as random, the task which is required is essentially a detection task in which the testee has to search for certain targets. The final test score (see the subsections on reliability) is based on the bar response times in the stationary part of the test. It is worth mentioning that a bar response time is the time spent on a bar, not the time spent on each individual target button. It is of utmost importance to notice that the variation in the bar response times (or better bar search times) across bars as well as across testees or across test administrations is entirely caused by the random displacements of the targets (the buttons to be clicked upon) in each bar. If the targets would have been positioned at fixed places, for example, at the positions 4, 10 and 16 of each bar having in all 18 positions, then one may expect that the time to complete each bar would have been more or less constant resulting in about the same total response time for each bar. This is especially true if the testee starts working rhythmically which is an efficient strategy to make it easy to continue work for a long time. In this respect, it is worthwhile to mention a study by Sanabria et al. (2011). This study explicitly examined the effect of rhythms on RT performance. Due to the random displacement of the target buttons, the individual target button times T_i will be dependent on the number of consecutive non-target buttons L_i that precede each target button (or the start button for the next bar). One such example would be $T_i = a + bL_i + E$, where a and b are subject dependent parameters, and E is a random error component

whose variability is also dependent on a subject parameter, say σ . From a psychological standpoint, the variable E is the most important because it represents both the mental stability and mental capacity of the subject. Depending on the value of b , the influence of L_i on T_i will be smaller or larger. Moreover, if the probability distribution of L_i is more symmetrical, the probability distribution of T_i will be more symmetrical as well. Although the probability distribution of T_i will be more complex than the exponential distribution, in the sequel to this paper, it will still be assumed that T_i has an approximate exponential distribution. This will especially be the case if the influence of L_i on T_i is small or if the distribution of L_i is strongly skewed to the right. The main reason for choosing the exponential distribution is the principle of parsimony. The logic behind the exponential probability distribution is rather simple (constant transition rate). A simple explanation is preferred over a complex one unless the data dictate otherwise.

The fixed condition of the test: the Erlang distribution

In the default version of the test, the number of target buttons in each bar equals three. This implies that the total time needed for a bar is the sum of four individual response times:

- 1 The time that elapses between the start button click and the first target button click.
- 2 The time that elapses between the first target button click and the second target button click.
- 3 The time that elapses between the second target button click and the third target button click.
- 4 The time that elapses between the third target button click and the start button click for the next bar.

It is assumed that each individual response time has an approximate exponential distribution. Except for the arguments given above, it could only be an approximation because the distributions of the first and fourth individual response times are different from the second and third individual response times.

If one would assume that the individual button click times are independent and exponentially distributed then a bar response time has an Erlang distribution with the shape parameter k being equal to the number of target buttons plus one. This is done, because one has to take into account the time that elapses between the last target button click and the start button click for the next bar. The number of search times, therefore, is not equal to the number of target buttons but equal to the number of target buttons plus one. For the default version of the test the number of target buttons is equal to three and, therefore,

$k = 4$. In summary, it is assumed, that in the case of the fixed condition of the test the probability distribution $f(t)$ of a response time T for a given test administration for a separate bar is given by the shifted Erlang distribution:

$$f(t) = \frac{\lambda^k (t-A)^{k-1} e^{-\lambda(t-A)}}{(k-1)!} \quad (1)$$

with expectation

$$\mu_T = A + \frac{k}{\lambda} \quad (2)$$

and variance

$$\sigma_T^2 = \frac{k}{\lambda^2} \quad (3)$$

From these two equations, it follows that

$$\sigma_T = \frac{\mu_T - A}{\sqrt{k}} \quad (4)$$

The theoretical mean μ_T can be estimated by the observed mean (m) of the response times in the stationary part of the time series, the theoretical standard deviation σ_T by the observed standard deviation (s) and the shift parameter A by the minimum of all the response times.

The random condition of the test: the Solleveld Erlang distribution

In the random condition of the test, the number of consecutive button clicks that have to be made for each bar may be different for each bar and randomly vary across bars. Consequently, the total time needed for a bar is no longer a shifted Erlang distribution because the number of button clicks to be made is not fixed but random. The question arises as to how, within a single test administration, the distribution would be, given the fact that the number of button clicks in a bar is random. This distribution is a mixture of a shifted Erlang distribution (with a shape parameter k) with some other discrete probability distribution for K . The shape of this distribution follows from the way the buttons that have to be clicked are distributed across bars. This distribution appears to be a so-called truncated Solleveld distribution that was developed specifically for this paper and is described in detail in the Appendix. Therefore, for a single test administration, the probability distribution of a bar response time is a compound probability distribution existing of a shifted Erlang distribution with parameter k , which in turn is distributed according to a truncated Solleveld distribution. The resulting distribution is marginalized over the truncated Solleveld distribution (i.e., summing the k out). This distribution is

said to be the distribution that results from combining the shifted Erlang distribution with the truncated Solleveld distribution. The distribution is called the Solleveld Erlang distribution.

The question arises as to how the probability distribution of the number of buttons to be clicked upon will appear. This problem can be solved by considering all possible dice sequences consisting of six possible different types of dice displayed at 18 positions such that no consecutive dice of the same type occur. This is exactly the same as all possible words (as defined in the Appendix) of length 18 in an alphabet of six symbols without repetitions (as defined in the Appendix). The number of words of length n in an alphabet of symbols without repetitions in which a distinguished element appears precisely k times is given in formula (A.3). This formula is used to define the so-called Solleveld distribution as follows:

$$P(K = k) = \frac{\sum_{i=0}^2 \binom{2}{i} \binom{n-k-1}{k-i} (j-1)^{k+1-i} (j-2)^{n+i-1-2k}}{j(j-1)^{n-1}} \quad (5)$$

where for n is even $k = 0, 1, 2, \dots, \frac{1}{2}n$ and for n is odd $k = 0, 1, 2, \dots, \frac{1}{2}(n+1)$ and $j \leq n$. The expectation of K , $E(K)$ is equal to n/j regardless of whether n is even or odd. It must also be considered, however, that no button sequences occur that contain no dice with the dots horizontally positioned in the middle. At least one dice of this type will always be present in a bar. Therefore, instead of the above defined Solleveld distribution, the truncated Solleveld distribution applies here:

$$P(K = k) = \frac{\sum_{i=0}^2 \binom{2}{i} \binom{n-k-1}{k-i} (j-1)^{k+1-i} (j-2)^{n+i-1-2k}}{j(j-1)^{n-1} - (j-1)(j-2)^{n-1}} \quad (6)$$

where for n is even $k = 1, 2, \dots, \frac{1}{2}n$ and for n is odd $k = 1, 2, \dots, \frac{1}{2}(n+1)$ and $j \leq n$. Since for the ACT, only the case $j = 6$ (six different buttons) and $n = 18$ (eighteen buttons in a bar) is of interest, from now on only the formulas for this case will be discussed. The expectation of K for $j = 6$ and $n = 18$ is as follows:

$$E(K | j = 6, n = 18) = \frac{1373291015625}{449173737283} = 3.057. \quad (7)$$

This value slightly differs from the expectation of the untruncated Solleveld distribution. In the fixed condition of the test, it was assumed that the density function f has an Erlang distribution with unknown discrete shape parameter k and rate parameter λ (see formula (1)):

$$f(t | k) = \frac{\lambda^k t^{k-1} e^{-\lambda t}}{(k-1)!} \quad \text{with } 0 \leq t$$

and g is a truncated Solleveld distribution $P(K = k | j = 6, n = 18)$ according to formula (6). The compound distribution has the following form:

$$h(t) = \sum_{k=1}^9 \frac{\lambda^{k+1} t^k e^{-\lambda t} P(K = k | j = 6, n = 18)}{k!} \quad (8)$$

where $P(K = k | j = 6, n = 18)$ refers to formula (6). Notice that k in $P(K = k)$ corresponds to with $k + 1$ in $f(t|k)$. This is done, because one has to take into account the time that elapses between the last target button click and the start button click (for the next bar). The person cannot know whether the last target button is indeed the last target button since he/she does not know how many target buttons there are in a bar. After the last target button, the person still has to search for other targets. The number of search times, therefore, is not equal to the number of target buttons but is equal to the number of target buttons plus one. The expectation $E(T) = \mu_T$ of T in formula (8) can be written as follows:

$$\mu_T = A + \frac{1822464752908}{449173737283} \frac{1}{\lambda} \quad (9)$$

Notice that the shift parameter A is included in the formula. The variance $Var(T) = \sigma_T^2$ can be written as follows:

$$\sigma_T^2 = \frac{1133838923256416857864726}{201757046264777504222089} \frac{1}{\lambda^2} \quad (10)$$

From these two equations it follows that

$$\sigma_T = \frac{\sqrt{235369212550315378519419610671399771382813018054}}{830344443898004373614116} \mu_T \quad (11)$$

The theoretical mean μ_T can be estimated by the observed mean (m) of the response times in the stationary part of the time series, the theoretical standard deviation σ_T by the observed standard deviation (s) and the shift parameter A by the minimum of all the response times.

Purpose of the study

Any method of scoring used in a psychological test should be theoretically based. If the obtained data are response times, which are usually considered random variables, at least the probability distribution of these response times should be known in advance to make it possible to derive a proper method of scoring the test. The ACT scoring could be based on the bar response times or, at a deeper level, on the individual button click times. Since

up till now scoring has been based on the bar response times our main focus will be on these response times. In this study, we will determine how the predictions, which were made according to the proposed probability distributions, compared to their observed counterparts. Two different samples were used: one in which all subjects performed the fixed colours version of the test and one in which all subjects performed the random positions version of the test. In the fixed colours sample, the outcome of the predictions for the Erlang distribution was checked, whereas in the random positions sample, it was the outcome of the predictions of the Solleveld Erlang distribution.

Method

The data are a matrix of response times consisting of N rows of test administrations and n columns of consecutive response times (t_{ij} , $i = 1, \dots, N$, $j = 1, \dots, n$). The N test administrations may consist of more than one test administration per subject. Various probability distributions have been proposed in the past to describe the observed frequency distribution of the response times. However, it is not always clear whether the proposed distribution applies across trials $j = 1, \dots, n$ or across test administrations $i = 1, \dots, N$. This paper focusses on the development and application of probability distributions for the response times obtained in a single test administration.

Participants

In this study, two different data sets were used. Firstly, to determine whether or not the Erlang distribution corresponded to the empirical findings, we used data obtained by the fixed version of the Attention Concentration Test. The chosen test parameters included colours for the task and fixed for the stimulus number. The subjects were Finnish primary school students who ranged in age from 8 to 12 years. The number of test administrations (cases) was equal to 362. Fifty-four participants completed the test several times (see Table 1), which provided a reliability estimate for the consecutive test rounds. The testing took place during January, February and December 2010 and January and February 2011. These tests were conducted to determine how the Attention Concentration Test works among different student groups during the primary school years.

Secondly, to determine if the Solleveld Erlang distribution corresponded to the empirical findings, we used data obtained by the random version of the Attention Concentration Test. The chosen test parameters included positions for the task and random, restricted for the stimulus number. The subjects ($N = 334$) were 16 years old. One hundred and

Table 1: Frequency distribution of the *number* of times the test was performed in the fixed condition of the test. The number of subjects that performed the test one time was equal to 149, the number of subjects which performed the test two times was equal to 54, etc.

<i>Number</i>	Male	Female	Total
1	82	67	149
2	24	30	54
3	13	3	16
4	3	1	4
5	0	7	7
6	1	0	1
Total	123	108	231

fifty-one subjects completed the test several times, which provided a reliability estimate for the consecutive test rounds. The testing took place during February, March, April and May in 2007. These tests were conducted to determine how attention measured by the Attention Concentration Test both correlates and contributes to scientific reasoning (Hotulainen et al., 2014).

The Attention Concentration Test (ACT)

The ACT primarily measures attention, specifically the level of concentration, as measured by variations in a series of reaction times. The computer-based ACT has the following options:

- Test-length: Beginner (25 bars), Intermediate (35 bars), Expert (45 bars)
- Difficulty: Easy (one target), Medium (two targets), Difficult (three targets)
- Task: Colours, Positions, Dice
- Instruction: Quickly vs. Relaxed
- Stimulus-number: Fixed vs. Random.

The first one of each category is the default option.

The fixed colours task

The default version of the test consists of a series of 25 bars of buttons representing colours. The following colours occur: blue, green, yellow, orange, red and purple. Each colour occurs three times. The colour buttons are displayed in random order. Figure 1 shows an example of a particular bar. The randomization was as follows. Each bar



Figure 1: Sample bar of the colours task of the Attention Concentration Test

consisted of 18 buttons. Each colour appeared randomly in the first six consecutive buttons. Each colour appeared randomly in the second six consecutive buttons. Each colour appeared randomly in the third six consecutive buttons. If two consecutive colours were the same, then a new bar was created with a new arrangement of colours. This was repeated until no two consecutive colours were the same. The instructions were as follows. Click on the buttons with the colour red. Work as quickly as possible, making no errors. Click on the colours in the order they are given in the colour bar starting from left to right. In order to go to the next colour bar, you must click on the Next button, which is located underneath the colour bar. You are not allowed to make errors. Errors include the following:

- Omissions: no click on a required button (miss)
- True errors: a click on a non-required button (false alarm)
- Wrong order: clicking the required buttons in the wrong sequence

If one of these errors is made, the test is stopped and the subject has to start all over again.

Reliability of the fixed colour task

In the fixed condition the test was scored in the following way. For each individual test administration a linear regression analysis was performed with the response time as dependent and the sequence number of the last target button as the independent variable. The analysis was based on the response times and numbers of the last target button of the bars 16 - 25. This was done in order to secure stationary response times. The Standard Error of the Estimate also known as the square root of the Mean Square Residual was

used as the final score. In this way the variability of the stationary response times, measured independently from the sequence number of the last target button, is used as an index of test performance. The smallest response time based on all bars 1 - 25 was used as an index of pure working time performance.

In the fixed color sample, two test administrations were available for 54 subjects. The subjects varied in age from 8 to 12 years with frequencies respectively equal to 2, 5, 15, 29 and 3. The test-retest correlation of the The Standard Error of the Estimate of the response times 16 - 25 was equal to 0.642. The test-retest correlation of the natural logarithm transformation was equal to 0.521. The correlations of the Standard Error of the Estimate (first and second test administrations) with age in years were equal to -0.495 and -0.456, respectively. The corresponding correlations of the natural logarithm transformations were equal to -0.427 and -0.355, respectively. The correlation of age in years and age in days for this sample was equal to 0.898. Hence, the reliability of age in years was equal to $0.898^2 = 0.806$. The test-retest correlations and the correlations with age in years (corrected for attenuation) showed that the test has both sufficient reliability and validity. The test-retest correlation of the minimum of the response times 1 - 25 was equal to 0.757 and of the logarithm of the minimum was equal to 0.777.

The random positions task

In addition to the default version of the test, which has a fixed number of target buttons in each bar, another version of the test was used in this study. In the other version, the number of target buttons was random across bars. The other version had the same options as the default version of the test, except the categories Task and Stimulus-number. In the Task category, this version had the following: Positions with as Stimulus-number: Random, restricted. Each bar consisted of the following types of dice:

- Two dots horizontal in the middle
- Two dots vertical in the middle
- Two dots horizontal upwards
- Two dots horizontal downwards
- Two dots vertical to the right
- Two dots vertical to the left



Figure 2: Sample bar of the positions task of the Attention Concentration Test

The subject has to click on the dice in which the dots are horizontal in the middle (i.e., the target dice). Figure 2 shows an example of a particular bar. The dice were displayed in random order. The randomization was as follows. Each bar consisted of 18 dice. For each button on a bar, one of the above types of dice was randomly chosen with a probability of 1/6 for each type and then taken as that particular button. If in the resulting bar two consecutive dice were the same or if the target dice did not occur at least once, then a new bar was created with a new arrangement of dice. This process was repeated until no two consecutive dice were the same, and the target dice occurred at least one time. Underneath each bar, a separate "Next" button was displayed. The instructions were as follows: Click on the dice in which the dots are horizontal in the middle. Work as quickly as possible, making no errors. Click on the dice in the order they are given in the dice bar starting from left to right. In order to go to the next dice bar, you must click on the Next button, which is located underneath the dice bar. Errors are not allowed. If an error is made, the test is failed and the subject has to start all over again.

Reliability of the random positions task

In the random condition the test was scored in the following way. For each individual test administration a multiple regression analysis was performed with the response time as dependent and with the number of target buttons in a bar and the sequence number of the last target button as the explanatory ones. The analysis was based on the data obtained for the bars 16 - 25. This was done in order to secure stationary response times. The Standard Error of the Estimate was used as the final score. In this way the variability of the stationary response times, measured independently from the number of target buttons and from the sequence number of the last target button, is used as an index of test performance. The smallest response time based on all bars 1 - 25 was used as an index of pure working time performance. In the random positions version of the test, a sample of two test administrations consisting of 156 subjects was used. The age of 150 subjects was equal to 16. One subject was 11 years old, one subject was 15 years old and four subjects 17 years. The testing of the subject with the age of 11 years took place in January 2007. The testing of all other subjects took place during the months April and May 2007. One subject showed extremely low values for the Standard Error of the Estimate (0.092 and 0.073), and three subjects showed extreme high values during the

first test administration (3.641, 3.152 and 3.151). These subjects were removed from the sample as well as the 11-year-old subject, which left 151 subjects. The test-retest correlation of the Standard Error of the Estimate of the response times 16 - 25 was equal to 0.488. The test-retest correlation of the natural logarithm transformation was equal to 0.395. The test-retest correlation of the minimum of the response times 1 - 25 was equal to 0.755. The test-retest correlation of the logarithm of the minimum was equal to 0.669. The subjects did not differ very much in age. It did not make sense to compute correlations with age because most subjects (147 out of 151) were of the same age (16 years).

Analysis

The primary motivation for this study was to discover whether it is feasible to use scoring methods that are based on bar response times. For this endeavour, a proper probability distribution for the bar response times ($T_i, i = 1, 2, \dots, n$) must be developed, such as that obtained in a single test administration. Note that the collection of various test administrations may consist of more than one test administration for each subject. The Erlang distribution was used to account for the response times obtained under the test condition Stimulus-number: fixed. The Erlang distribution does not apply in the condition Stimulus-number: random, restricted because in that condition, the number of targets in a bar is random. The Solleveld Erlang distribution was used for the response times obtained under the test condition Stimulus-number: random, restricted. This distribution takes into account the random variation of the targets within each bar. Because the number n of stationary response times is very low ($n = 10$, see below), it does not make sense to perform a separate goodness of fit test for each test administration, such as by comparing the observed frequency distribution with the expected frequency distribution. Instead, an overall linear regression analysis of the standard deviation of the response times in each test administration was performed on the mean of the response times corrected for shift. A separate analysis was conducted for the fixed condition of the test as well as for the random, restricted condition. The observed response times may show some trends, especially at the beginning of the test. In the long run, the response times become stationary. Because the probability distributions require stationary and independent response times, only the response times in the stationary part of the time series were used. Before the actual regression analyses were performed, a trend analysis was conducted based on all response times ($n = 25$), and the outliers were defined on face value by examining the scatter plots given in the respective figures.

Results

The fixed colours version of ACT

The bar response times usually showed some trends in which lower response times were at the beginning of the test, and then gradually reached a stationary stage. In checking underlying probability distributions, the response times should be stationary. Therefore, before the probability distribution of interest was checked, first a trend analysis was performed on the bar response times.

Trend analysis

A plot of the means of the 25 consecutive reaction times is shown in Figure 3. Figure 3 shows a clear increasing long-term trend. After the first three reaction times, the means of the reaction times gradually increased and, in the long run, became stationary. The increase in the reaction times, which gradually reached the stationary state, could be explained by assuming that at the beginning of the task, the subjects generally worked much faster than they did later in the test when they grew accustomed to the task. The increasing curve should not be interpreted in terms of fatigue or learning. Gradually reaching one's 'natural' speed leads to increasing curves, which are marginally decreasing, whereas fatigue leads to increasing curves, which are marginally increasing (see van der Ven, 2001). Learning leads to decreasing curves. Figure 3 shows that the stationary state of the reaction times was reached at about the 16th reaction time. In order to check whether this was indeed the case, a repeated measures analysis of variance was performed with the consecutive reaction times 16 - 25 as the within-subjects factor. Neither the repeated-measures linear polynomial contrast ($p = 0.731$) nor any of the higher order polynomial contrasts was significant at the 5% level. It was decided from the beginning on to check the repeated-measures linear polynomial contrasts with numbers of consecutive response times decreasing in blocks of five. The first check was based on the consecutive response times 1 - 25, the second on the consecutive response times 6 - 25, the third on the consecutive response times 11 - 25 and so on. With a repeated measures analysis of variance with the consecutive reaction times 11 - 25 as the within-subjects factor the linear polynomial contrast was still significant at the 5%-level ($p = 0.033$). It was therefore decided to base all further analyses on the consecutive response times 16 - 25.

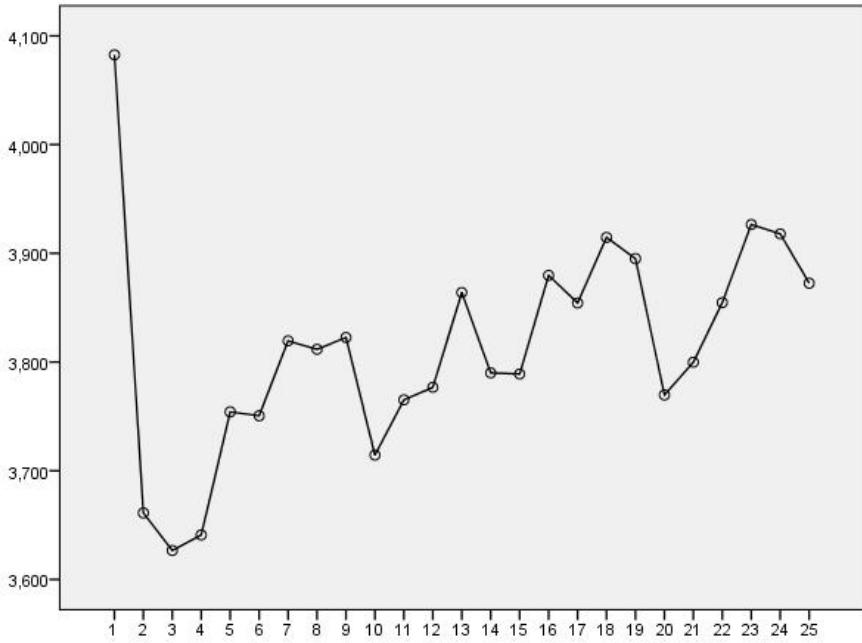


Figure 3: Average bar response times in seconds for the fixed condition of the test

Checking the Erlang distribution

If the Erlang distribution holds, then according to formula (4), the linear regression of s on $m - \hat{A}$ should be linear with an intercept approximately equal to 0 and a slope approximately equal to $1/\sqrt{k}$. The stationary mean m and the stationary standard deviation s were computed from the RTs 16 - 25. The minimum as an estimate for A was computed from the RTs 1 - 25. The scatter plot of the regression of s on $m - \hat{A}$ is given in Figure 4. The inspection of the scatter plot showed the existence of six extreme values of s . These cases were eliminated from further analyses. A linear regression was conducted with $m - \hat{A}$ as the independent variable and s as the dependent variable. The intercept a was equal to -0.047 ($N = 356$) and the slope b to 0.713 ($N = 356$). The intercept a did not deviate significantly from 0 at the 5%-level ($t = -1.746$, $p = 0.082$).

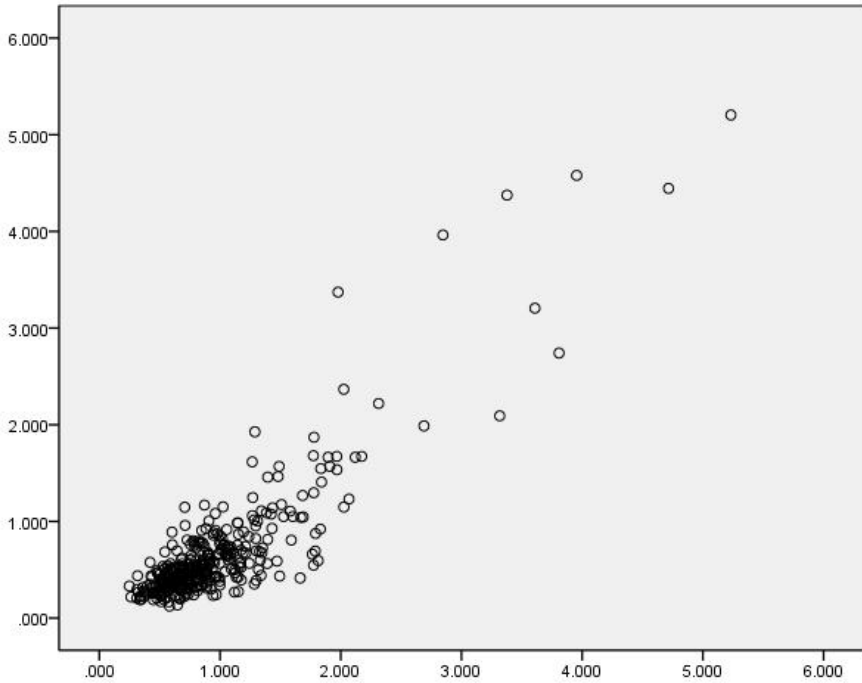


Figure 4: The regression plot of the standard deviation of the response times 16-25 (vertical axis) on the difference between the mean of the response times 16-25 and the minimum of the response times 1-25 (horizontal axis) for the fixed condition of the test

According to formula (4), the predicted slope \tilde{b} would be $1/\sqrt{k}$. Therefore, for $k = 4$ one would obtain $\tilde{b} = 1/\sqrt{4} = 0.500$. The difference between the observed value of the slope 0.713 and its theoretical value 0.500 was significant. The significance test was performed as follows. Firstly, the predicted standard deviation was computed by taking $\beta(m - \hat{A})$ with $\beta = 1/\sqrt{4}$. Next, the residual was computed by taking $s - \beta(m - \hat{A})$. Finally, a simple linear regression analysis was performed of $s - \beta(m - \hat{A})$ on $(m - \hat{A})$. If the predicted slope of the regression line of s on $(m - \hat{A})$ is equal to β , then it could be expected that the slope of the linear regression of $s - \beta(m - \hat{A})$ on $(m - \hat{A})$ would not significantly differ from zero. The observed slope was equal to 0.213 ($F = 66.546$,

$df_1 = 1, df_2 = 354, p = 0.000$). This result is not in favour of the Erlang distribution with $k = 4$. Therefore, the explanation of the response times by an Erlang distribution with $k = 4$ must be rejected.

The number of non-targets before the first target button and after the last target button is half the number of non-targets between consecutive target buttons. Therefore, the influence of these numbers (before the first target and after the last) may be less. Moreover, as soon as a new bar appears, knowing that only three targets will occur, the subject may see at a glance the location of the first target-button making it not really necessary to check all non-targets preceding. In addition, as soon as the last target button has been clicked upon, and again knowing how many targets occur in each bar, the subject will click on the start button for the next bar without looking for other target buttons after the third target button. These factors may reduce the influence of the time preceding the first target button click and the time following the last target button click on the total bar time. Therefore, instead of taking $k = 4$ one could also take $k = 2$. For $k = 2$ one would obtain $\tilde{b} = 1/\sqrt{k} = 0.707$. With $k = 2$ the difference between the observed value of the slope 0.713 and its theoretical value 0.707 is less pregnant. With the same method of significance testing as before the difference was not significant. The observed slope of the linear regression of $s - \beta(m - \hat{A})$ on $(m - \hat{A})$ was equal to 0.006 ($F = 0.046, df_1 = 1, df_2 = 354, p = 0.831$). This result is in favour of the Erlang distribution with $k = 2$.

The random positions version of the ACT

Trend analysis

A plot of the means of the 25 consecutive reaction times is shown in Figure 5. Except for the first reaction time, Figure 5 shows a clearly increasing long term trend. An analysis of variance was performed using a repeated measures design with one within-subjects factor with 24 levels: the consecutive reaction times 2 - 25. The repeated measures design is also known as a within-subject design. The right tail probabilities for the repeated-measures polynomial contrast tests for linearity and quadraticity were both equal to 0.000 with $N = 334$. The right tail probability for cubicity was equal to 0.017. Higher order polynomial contrasts were not significant at the 5% level. After the first reaction time, the means of the reaction times gradually increased and gradually became stationary. With the exception of the first average response time, the curve showed increasing average response times that were marginally decreasing. Note that fatigue would lead to marginally increasing curves. Figure 5 also shows that the stationary state of the reaction times was reached at about the 11th reaction time. In order to

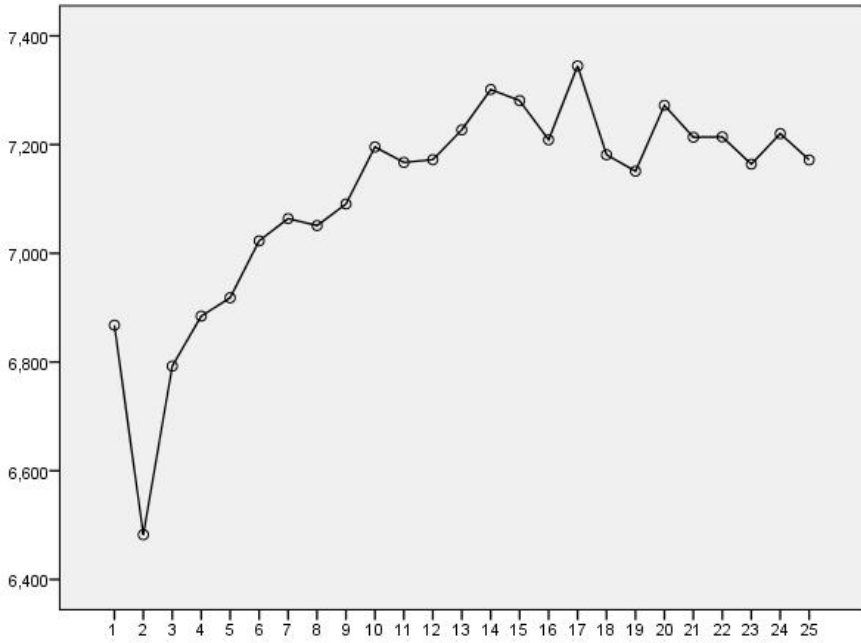


Figure 5: Average bar response times in seconds for the random condition of the test

check whether this was indeed the case, a repeated measures analysis of variance was performed with the consecutive reaction times 11 - 25 as the within-subjects factor. Neither the repeated-measures linear polynomial contrast ($p = 0.652$) nor any of the higher order polynomial contrasts was significant at the 5% level. It was again decided from the beginning on to check the repeated-measures linear polynomial contrasts with numbers of consecutive response times decreasing in blocks of five. The first check was based on the consecutive response times 1 - 25, the second on the consecutive response times 6 - 25, the third on the consecutive response times 11- 25 and so on. With a repeated measures analysis of variance with the consecutive reaction times 6 - 25 as the within-subjects factor the linear polynomial contrast was still significant at the 5%-level ($p = 0.035$). The quadratic contrast was also significant ($p = 0.003$) It was therefore decided to base all further analyses on the the consecutive response times 11 - 25.

Checking the Solleveld Erlang distribution

To test if the Solleveld Erlang distribution held then, according to formula (11) the linear regression of s on $m - \hat{A}$ should be linear with an intercept approximately equal to 0 and a slope β approximately equal to

$$\beta = \frac{\sqrt{235369212550315378519419610671399771382813018054}}{830344443898004373614116} \approx 0.584$$

The stationary mean m and the stationary standard deviation s were computed from the RTs 11 - 25. The minimum as an estimate for A was computed from the RTs 1 - 25. The scatter plot of the regression of s on $m - \hat{A}$ is shown in Figure 6. The scatter plot shows the existence of one extreme value of s . This case was left out from further analyses. A linear regression was conducted with $m - \hat{A}$ as the independent variable and s as the dependent variable. The intercept a was equal to 0.055 ($N = 333$) and the slope b to 0.549 ($N = 333$). The intercept a did not deviate significantly from 0 at the 5%-level ($t = 1.100$, $p = 0.272$). With the intercept fixed on 0, a slope equal to 0.575 was obtained. Note that this value was very close to the expected value according to the Solleveld Erlang distribution, which is equal to 0.584. The difference between the observed value of the slope 0.549 and its theoretical value was not significant. The significance test was performed in the same way as for the fixed condition of the test. The observed slope of the linear regression of $s - \beta(m - \hat{A})$ on $(m - \hat{A})$ was equal to -0.009 ($F = 0.741$, $df_1 = 1$, $df_2 = 331$, $p = 0.390$). This result supports the Solleveld Erlang distribution.

Discussion

The aim of this study was to determine whether the proposed RT distributions held with the empirical data. The Erlang distribution was proposed for the fixed condition of the test, and the Solleveld Erlang distribution was proposed for the random condition of the test. The Erlang distribution is the probability distribution of a sum of k , independent random variables, $k = 1, 2, \dots$, each of which has an exponential distribution. The Solleveld Erlang distribution is the probability distribution of a sum of K independent random variables each of which has an exponential distribution, where K is a random variable having a Solleveld distribution. In all these cases, the validity of the distribution was checked by applying a regression analysis of the standard deviation of the stationary response times on the mean corrected for shift. In the case of the fixed condition of the test the number of consecutive arrays containing only non-target buttons was equal to four, therefore, k was taken as equal to 4. The Erlang distribution with $k = 4$, however,

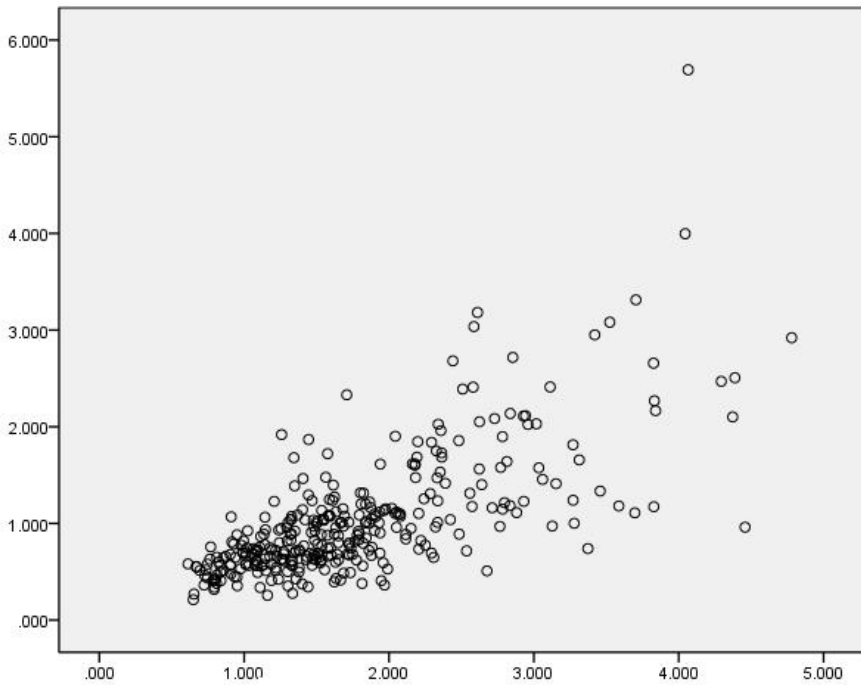


Figure 6: The regression plot of the standard deviation of the response times 16-25 (vertical axis) on the difference between the mean of the response times 16-25 and the minimum of the response times 1-25 (horizontal axis) for the random condition of the test

had to be rejected. This was not the case for $k = 2$. The main reason for taking $k = 2$ was, that the number of non-target buttons in the array preceding the first target button and the array following the last target button was only half of the number of non-target buttons in arrays between consecutive target buttons. Therefore, it seemed justified to assume that the effect of these beginning and end arrays on the bar response times could be neglected. In the case of the random condition of the test the Solleveld Erlang distribution was not rejected.

The statistical analysis in this study was entirely based on the bar response times. The bar response times were assumed to be dependent on only one latent variable, which

was further assumed to be exponentially distributed. This is an extremely simple state of affairs. Especially, if one is aware of the fact, that the bar response times themselves are dependent on the individual button click times, i.e. the times which elapse between the consecutive button clicks (start buttons and target buttons). It was argued that, because of the random displacement of the target buttons, the individual target button times T_i are dependent on the number of consecutive non-target buttons L_i preceding the target button (or start button for the next bar). Index i represents the sequence number of a target button (or start button for the next bar) in a bar. It was argued that the simplest way to put this would be to say $T_i = a + bL_i + E$, where a and b are subject dependent parameters and E is a random error component whose variability is also dependent on a subject parameter, say σ . Notice, that the component E is not just an 'error' term, but the most important component as it determines the variability in the bar response times. This setup, finally, means that, for each individual subject, the variability of T_i and, with it, also the variability of the bar response times is dependent on three subject parameters a , b and σ . In order to make it possible to test this simple model on its validity one should have the disposal of individual button click times instead of the bar response times such as used in this study. With individual button click times available, it would also be possible to make an attempt to discover the shape of the probability distribution of E .

Validation studies of self-paced continuous performance tasks often use the coefficient of variation c_v as a measure of response speed variability. This is done in addition to, or instead of the standard deviation s . However, if the probability distribution, which is used to explain the observed response times, contains a shift parameter, then the coefficient of variation c_v as well as the mean m of the observed response times both contain the shift parameter. This is, for example, quite clear if the shifted Erlang distribution (see formula (1)) is used as the explanatory underlying probability distribution. The theoretical mean is given by $A + k/\lambda$ (see formula (2)) and the theoretical variance by k/λ^2 (see formula (3)). Therefore, the theoretical coefficient of variation is given by $\sqrt{k\lambda}/(A\lambda + k)$. Actually, the coefficient of variation would be a bad measure as in the case of the shifted Erlang distribution it involves all the unknown parameters concerned. If one would have at his disposal an unbiased estimator of A , say \hat{A} , then, instead of the coefficient of variation, one could use $s^2/(m - \hat{A})$. For the shifted Erlang distribution this would be a moment estimator for $1/\lambda$. This measure could then be called *efficiency of work* and the measure $1/\hat{A}$ could be called *speed of work*.

References

- Binet, A. (1900). Attention et adaptation [Attention and adaptation], *L'Annee Psychologique*, 6, 248-404.
- Hotulainen, R., Thuneberg, H., Hautamki, J., & Vainikainen, M.-P. (2014). Measured attention in prolonged overlearned response tasks and its correlation to high-level scientific reasoning and school achievement. *Psychological Test and Assessment Modeling*, 56, 237-254.
- Huiskamp J., & de Mare, M. (1947). Enige opmerkingen over de Wiersma-Bourdon test [Somere remarks about the Wiersma-Bourdon test]. *Nederlands Tijdschrift voor de Psychologie [Dutch Journal of Psychology]*, 2, 75-78.
- Hylan, J. P. (1898). The fluctuation of attention. *The Psychological Review*. Series of Monograph Supplements. Vol. II, No. 2 (Whole No. 6). New York: The MacMillan Company.
- Jensen, A. R. (1982). Reaction time and psychometric g. In H. J. Eysenck (Ed.), *A model for intelligence* , (pp. 93-132). New York: Springer-Verlag.
- Klauer, K. C. (1993). *Belastung und Entlastung beim Problemlösen. Eine Theorie des deklarativen Vereinfachens.* [Cognitive load in problem-solving. Theory of declarative decision-making]. Göttingen: Hogrefe.
- Larson, G. E., & Alderton, D. L. (1990). Reaction time variability and intelligence: A "worst performance" analysis of individual differences. *Intelligence*, 14 (3), 309-325.
- Sanabria, D., Capizzi, M., & Correa, A. (2011). Rhythms That Speed You Up. *Journal of Experimental Psychology: Human Perception and Performance*, 37 (1), 236-244.
- Smit, J. C., & van der Ven, A. H. G. S. (1995). Inhibition in speed and concentration tests: The Poisson inhibition model. *Journal of Mathematical Psychology*, 39, 265-273.
- Spearman, C. (1927). *The abilities of man*. London: MacMillan.
- Steinborn, M. B., Flehming, H. C., Westhoff, K., & Langner, R. (2008). Predicting school achievement from self-paced continuous performance: Examining the contributions of response speed, accuracy, and response speed variability. *Psychology Science Quarterly*, 50, 613-634.

- Vos, P. G. (1992). *De Bourdon concentratietest voor kinderen* [The Bourdon concentration test for children]. Lisse: Swetz en Zeitlinger.
- van der Ven, A. H. G. S., Smit, J. C., & Jansen, R. W. T. L. (1989). Inhibition in prolonged work tasks. *Applied Psychological Measurement*, 13, 177-191.
- van der Ven, A. H. G. S. (2001). A theoretical foundation of speed and concentration tests. In F. Columbus (Ed.), *Advances in psychology research IV* (pp. 315-354). Huntington: Nova Science Publishers, Inc.

Appendix: Counting words without repetitions

by Maarten M.S. Solleveld¹

Consider an alphabet with j symbols, say $1, 2, \dots, j$. By a word of length n we shall mean a string of n symbols from this alphabet. These n symbols need not be different, so for each position we have precisely j possibilities. The number of words of length n is j^n .

We say that a word contains a repetition if there are two adjacent positions with the same symbol. We do not consider one symbol in two non-adjacent positions a repetition. For example

$$\begin{aligned} (23314) &: \text{repetition,} \\ (23143) &: \text{no repetition.} \end{aligned} \tag{A.1}$$

To construct an arbitrary word without repetitions, we work from left to right and fill the first position with any symbol from the alphabet. For each subsequent position, one symbol is excluded, so there are $j - 1$ choices. Hence, the number of words of length n without repetitions equals

$$j(j-1)^{n-1}. \tag{A.2}$$

From now on, one symbol will play a special role, namely the symbol 1. We are interested in words without repetitions, which contain precisely k ones, for a given k . The goal of this Appendix is to explain how such words can be counted. Our argument works only for words of lengths $n > 3$, so we assume that throughout.

Step 1. Suppose that the k positions containing the ones are fixed. We want to calculate the number of possible words.

¹Institute for Mathematics, Astrophysics and Particle Physics, Faculty of Science, Radboud University Nijmegen, P.O. Box 9010 (Heyendaalseweg 135), 6500 GL Nijmegen, The Netherlands

When we remove the ones, we are left with several shorter words without repetitions from the alphabet $\{2, 3, \dots, j\}$. For example,

$$(21341231) \mapsto (2), (34), (23).$$

For each of these shorter words, the number of possibilities is determined by (1.2), but with $j - 1$ instead of j and a smaller n .

Let us call the number of subwords without ones d and denote their lengths by n_1, n_2, \dots, n_d . Then

$$n_1 + n_2 + \dots + n_d = n - k.$$

We can choose the d subwords independently. In total there are

$$(j-1)(j-2)^{n_1-1}(j-1)(j-2)^{n_2-1} \dots (j-1)(j-2)^{n_d-1} =$$

$$(j-1)^d (j-2)^{n_1+n_2+\dots+n_d-d} = (j-1)^d (j-2)^{n-k-d}$$

possibilities. Note that this does not depend on the n_i , only on j, n, k and d .

Step 2. Determine how many placements of the ones cut the word in precisely d subwords.

Since there has to be at least one spot between any two ones, the number d depends only on whether the first and/or the last character of our word is a one. There are four cases to consider:

– Case a: Neither the first nor the last character is a one.

Here $d = k + 1$. We must count the number of ways we can select k positions from the middle $n - 2$ positions, such that no two adjacent spots are chosen. We claim that this equals

$$\binom{(n-2)-k+1}{k} = \binom{n-k-1}{k}.$$

Suppose we selected the positions m_1, m_2, \dots, m_k from $\{2, 3, \dots, n-1\}$, with $m_{i+1} > m_i + 1$. To this we associate the k integers $m_1 - 1, m_2 - 2, \dots, m_k - k$. These are all different and belong to $\{1, 2, \dots, n - k - 1\}$. Conversely, given any k elements m'_i of $\{1, 2, \dots, n - k - 1\}$, we associate to them the positions $m'_1 + 1, m'_2 + 2, \dots, m'_k + k$ in $\{2, 3, \dots, n - 1\}$. Then $m'_{i+1} + (i + 1) > (m'_i + i) + 1$.

This provides a bijection between the selections that we wish to count and selections of k elements from a set of $n - k - 1$ elements.

- Case b: there is a one at the first position, but not in the last position.

Then $d = k$. Since some other symbol occupies position 2, the remaining $k - 1$ ones are distributed over the $n - 3$ positions $\{3, 4, \dots, n - 1\}$, without repetitions. With the same method as used in case a we see that the number of ways to arrange this is

$$\binom{(n-3) - (k-1) + 1}{k-1} = \binom{n-k-1}{k-1}.$$

- Case c: there is a one at the last position, but not at the first.

Again $d = k$. By left–right symmetry, the number of possibilities is the same as in case b, that is, $\binom{n-k-1}{k-1}$.

- Case d: both the first and the last position are occupied by ones.

Here $d = k - 1$. The assumption entails that the positions 2 and $n - 1$ must be filled with other symbols. We still have $k - 2$ remaining occurrences of 1, which must be chosen from $\{3, 4, \dots, n - 2\}$ but not in adjacent positions. Since $n \geq 4$, this set contains exactly $n - 4$ elements. The argument for case a also applies here, and shows that the number of possibilities is

$$\binom{(n-4) - (k-2) + 1}{k-2} = \binom{n-k-1}{k-2}.$$

Step 3. Combine the previous calculations.

For each d we take the number of arrangements of ones which give rise to this number of subwords without ones, as obtained in Step 2, and we multiply it with the number of ways to fill in the subwords from Step 1. The result is the following:

$$\begin{aligned} & \binom{n-k-1}{k} (j-1)^{k+1} (j-2)^{n-2k-1} \\ + & 2 \binom{n-k-1}{k-1} (j-1)^k (j-2)^{n-2k} \\ + & \binom{n-k-1}{k-2} (j-1)^{k-1} (j-2)^{n-2k+1} \end{aligned}$$

which is equal to:

$$\sum_{i=0}^2 \binom{2}{i} \binom{n-k-1}{k-i} (j-1)^{k+1-i} (j-2)^{n+i-1-2k} \quad (\text{A.3})$$

This is the number of words without repetitions, from an alphabet of j symbols, of length n , in which a distinguished element appears precisely k times.

Note that the second term of the sum (1.3) is zero if $k = 0$, then

$$\binom{n-k-1}{k-1} = \binom{n-1}{-1} = 0.$$

Similarly, the third term of (1.3) is zero if $k = 0$ or $k = 1$. This fits because the second term corresponds to cases b and c in Step 2, which obviously cannot occur if $k = 0$. The third term of the sum comes from case c, which only occurs if $k \geq 2$.

When $n \leq 3$, the above argument fails because some positions considered as different would actually coincide. However, it is easy to work out those cases out by hand by using the same ideas. For completeness, we include the following results:

n	k	number of possibilities
1	0	$j - 1$
1	1	1
2	0	$(j - 1)(j - 2)$
2	1	$2(j - 1)$
3	0	$(j - 1)(j - 2)^2$
3	1	$(j - 1)^2 + 2(j - 1)(j - 2) = (j - 1)(3j - 5)$
3	2	$j - 1$

Remark. It seems to be much more difficult to count words without repetitions in which two distinguished elements appear a given number of times. Then the above steps are insufficient, and it is not clear how it can be counted without considering a large number of cases.