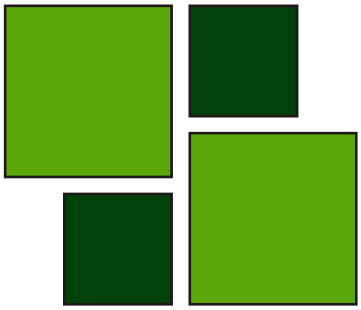

WWW.JAQM.RO

JOURNAL OF APPLIED QUANTITATIVE METHODS

Quantitative Methods Inquires

Vol. 6
No. 3
Fall
2011

ISSN 1842–4562

JAQM Editorial Board

JAQM Advisory Board

JOURNAL OF APPLIED QUANTITATIVE METHODS

Contents

JOURNAL OF APPLIED QUANTITATIVE METHODS

Quantitative Methods Inquires

COMPARING DISTRIBUTIONS: THE TWO-SAMPLE ANDERSON-DARLING TEST AS AN ALTERNATIVE TO THE KOLMOGOROV-SMIRNOFF TEST

Sonja ENGMANN [1]
PhD, University of Montreal, Canada

Denis COUSINEAU[2]
PhD, University Professor,
University of Ottawa, Ottawa, Canada

E-mail: denis.cousineau@uOttawa.ca

Abstract:
This paper introduces the two-sample Anderson-Darling (AD) test of goodness of fit as a tool for comparing distributions, response time distributions in particular. We discuss the problematic use of pooling response times across participants, and alternative tests of distributions, the most common being the Kolmogorov-Smirnoff (KS) test. We compare the KS test and the AD test, presenting conclusive evidence that the AD test is more powerful: when comparing two distributions that vary (1) in shift only, (2) in scale only, (3) in symmetry only, or (4) that have the same mean and standard deviation but differ on the tail ends only, the AD test proves to detect differences better than the KS test. In addition, the AD test has a type I error rate corresponding to alpha *whereas the KS test is overly conservative. Finally, the AD test requires less data than the KS test to reach sufficient statistical power.*

Key words: Anderson-Darling test; Kolmogorov-Smirnoff test; Comparing distributions

Introduction

The motivation for this article lies in the authors' own research on redundancy gain (Miller, 1982): we investigate response time (RT) distributions in an object recognition task, varying the number of redundant attributes identifying an object as a target (Engmann & Cousineau, submitted). We analyze each participant's RTs individually, and therefore needed a test that would allow analysis of a whole distribution, not just the mean and variance. We wanted a test that is sensitive to changes in shape and asymmetry. After trying different goodness-of fit tests, we finally settled on the Anderson-Darling test, a powerful tool for comparing data distributions. In this paper we wish to introduce the two-sample version of the Anderson-Darling (AD) test and compare its power to the Kolmogorov-Smirnoff (KS) test.

The AD test is commonly used in engineering, but little known in Cognitive Psychology, despite its advantages for this field. This test is especially useful if there is not a lot of data available in the samples to be compared, and when the analysis should extend beyond distributions' means, taking into account differences in shape and variability as well as the mean of the given distributions. The AD test is non-parametric and can be applied to Normal, Weibull, and other types of distributions (Isaic-Maniu, 1983, Cousineau, Brown & Heathcote, 2004, Gumbel, 1958, Galambos, 1978). It is especially useful to analyze response time distributions, as it allows a participant-by-participant analysis.

Why combining response time distributions across participants is problematic

When comparing response time (RT) distributions for different experimental conditions, it can be quite difficult to obtain a sufficient amount of data in each condition for a reliable analysis. There is a trade-off between the time participants take for a given experiment and the amount of data per condition. Combining the response times of several participants seems to be, at first glance, an elegant solution to avoid this trade-off. However, on closer inspection, combining RT distributions presents several difficulties.

The most intuitive solution, simply pooling all RTs from all participants together per condition, would produce uninterpretable distributions due to inter-participant variability: such RT distributions would not only be influenced by the characteristics of the experimental condition under which they are produced, but also by individual differences. Participants can have faster or slower motor reactions, or object recognition speed – the possibilities to produce variance in RT distributions are endless – such that variance between participants will be larger than variance due to experimental manipulation. Therefore, simple pooling of different RT distributions will flatten the shape of the final distribution, or, if there are not many participants, lead to a bi- or multimodal distribution.

A technique to avoid some of these problems was proposed by Vincent (1912; see also Rouder & Speckman 2004). The so-called Vincentizing is the most popular technique to combine response time distributions. It involves dividing each distribution into a certain number of quantiles, and then averaging the nth quantiles of each distribution. The advantage of using this technique is that the resulting "average" RT distribution takes into consideration the relative position of each response time in relation to the other RTs of a specific participant, i.e. minimal RTs are averaged with other minimal RTs; RTs at the peak of each participant's distribution are averaged with other peaks; etc. This avoids a flattening or multi-modality of the Vincentized distribution.

However, Vincentizing distorts the shape and symmetry of individual distributions (Thomas & Ross, 1980). If an RT distribution reflects one or more underlying processes that contribute to the RT, then this information is essential for analysis. A Vincentized distribution tends towards normality, whereas asymmetry is a universal finding in RT empirical data (Logan, 1992; Rouder, Lu, Speckman, Sun & Jiang, 2005). Possibly relevant information about a RT distribution, such as its degree of symmetry, gets lost when Vincentizing.

Vincentizing is the best technique of combining RT distributions available right now. However, even Vincentizing does not render an unbiased and exact analysis of RT distributions, and research for a better method is in progress, but has not been conclusive so far (Lacouture & Cousineau, in press). Therefore, we need to consider methods available for participant-by-participant analysis.

Different methods of comparing distributions participant-by-participant

The most common methods of comparing two or several distributions, the t-test or the ANOVA, render a judgment of goodness of fit based on the mean and variance of distributions under comparison. They do not take shape and symmetry into account, which is not specific enough in a lot of cases, for reasons mentions in the previous section. Also, both tests are parametric, expecting a normal distribution, whereas RTs have a shape close to the Weibull or the Lognormal distribution.

When investigating redundant target recognition RTs, several authors used multiple t-tests on quantiles (Miller, 1982; Mordkoff & Yantis, 1991, 1993, among others). Quantiles (e. g. the 5^{th} percent quantiles) are computed for each participant in the two conditions whose distributions are to be compared. These quantiles are then tested for equality using a t-test. This procedure is replicated for all quantiles at given intervals (e. g. the 10^{th}, the 15^{th}, etc. percent). This method allows an estimate of where RT distributions of all participants differ significantly. It keeps individual participants' data separate, and analyses more than distribution means.

However, sample size for each t-test is only as large as the number of participants in an experiment; therefore statistical power may not be sufficient, especially if the effect size is not very large to begin with. Additionally, between-participant variability might be larger than between-condition differences. Finally, the data at one time point are highly correlated with the data at the previous and following time point, influencing the probability of a type I error rate.

There are several types of non-parametric or distribution-free (they neither depend on the specific form, nor on the value of certain parameters in the population distribution; Massey, 1951) goodness of fit tests that either test if a sample comes from a given theoretical distribution, or if two samples come from the same underlying distribution. The most well-known in psychology, although used more frequently as a test of independence than goodness of fit, is the Pearson's Chi square (χ^2) test (Chernoff & Lehmann, 1954). The χ^2 test operates on binned frequency distributions, not on probability distributions, and does not give precise results when bin size is too narrow. It is therefore less adapted and less powerful than other tests for comparison of distributions, such as the Kolmogorov-Smirnoff, Cramer-von Mises, Kuiper, Watson or Anderson-Darling test (Stephens, 1974). All of the above tests have more or less the same underlying structure, or are adaptations of one another for different sample sizes or situations, some being more powerful for detecting changes in mean, others in variance (Stephens, 1974).

The Kolmogorov-Smirnoff (KS) test is the most well-known of these tests, and the most commonly used in psychology. The KS test's statistical power is greater than that of the χ^2-test, it requires less computation, and unlike the latter, it does not lose information by binning, as it treats individual data separately (Massey, 1951; Lilliefors, 1967). However, it is applicable neither for discrete distributions, nor in cases where not all parameters of a theoretical distribution are known and therefore, they have to be estimated from the sample itself.

In this article, we will concentrate on a comparison of the Kolmogorov-Smirnoff (KS) and the Anderson-Darling (AD) test. The former test is already commonly used in the field of psychology, and both are non-parametric, distribution-free, do not require normality, and are best adapted to the context of RT distribution analysis.

Comparison of Kolmogorov-Smirnoff and Anderson-Darling tests

Both the KS and the AD test are based on the cumulative probability distribution of data. They are both based on calculating the distance between distributions at each unit of the scale (i.e. time points for RT distributions).

Kolmogorov-Smirnoff Test

The Kolmogorov-Smirnoff (KS) test was first introduced by Kolmogorov (1933, 1941) and Smirnoff (1939) as a test of the distance or deviation of empirical distributions from a postulated theoretical distribution. The KS statistic for a given theoretical cumulative distribution *F(x)* is

$$KS_n = \sqrt{n}\sup_x |F_n(x) - F(x)| \quad (1)$$

where *F(x)* is the theoretical cumulative distribution value at x, and $F_n(x)$ is the empirical cumulative distribution value for a sample size of n. The null hypothesis that $F_n(x)$ comes from the underlying distribution *F(x)* is rejected if KS_n is larger than the critical value KS_α at a given α (for a table of critical values for different sample sizes see Massey, 1951; less conservative critical values exist if the test distribution is the normal distribution, Lilliefors, 1967, or the exponential distribution, Lilliefors, 1969). This means that a band with a height of KS_α is drawn on both sides of the theoretical distribution, and if the empirical distribution falls outside that band at any given point, the null hypothesis is rejected. The KS-statistic is sometimes abbreviated as D-statistic. For reasons of clarity we will use the former term throughout this article.

The two-sample version of the KS test generalizes to

$$KS_{nn'} = \sqrt{\frac{nn'}{n+n'}}\sup_x |F_n(x) - F_{n'}(x)| \quad (2)$$

where $F_n(x)$ and $F_{n'}(x)$ are two empirical cumulative distribution values at time point *x*, based on data sets of size *n* and *n'* respectively. The null hypothesis that $F_n(x)$ and $F_{n'}(x)$ come from the same underlying distribution is rejected if $KS_{n\,n'}$ is larger than the critical value KS_α at a given α (for a table of critical values for the two-sample KS test, see Massey, 1951).

The main advantage of the KS test is its sensitivity to the shape of a distribution because it can detect differences everywhere along the scale (Darling, 1957). Also, it is applicable and dependable even for small sample sizes (Lilliefors, 1967). Therefore, a KS test is advised in the following experimental situations: (1) when distribution means or medians are similar but differences in variance or symmetry are suspected; (2) when sample sizes are small; (3) when differences between distributions are suspected to affect only the upper or lower end of distributions; (4) when the shift between two distributions is hypothesized to be small but systematic; or (5) when two samples are of unequal size.

The KS test is fairly well known in the field of psychology, and has been used for a number of different experimental contexts other than a comparison of response times, such as a comparison of circadian rhythm (Pandit, 2004), an evaluation of exam performance (Rodriguez, Campos-Sepulveda, Vidrio, Contreras & Valenzuela, 2002), or a comparison of economic decision-making (Eckel & Grossman, 1998).

Initially, the authors also used the KS test to compare the response times of participants in an object recognition task where objects could be defined by one, two or three target attributes (Engmann & Cousineau, submitted). However, we began looking for an alternative for the following reasons. First, participants were faster at recognizing objects defined by several target attributes, but the effect was very small. Second, we wanted to compare our data to a model which made certain assumptions about minimal response times, as well as scale and symmetry of response time distributions. We therefore needed a test that would detect small differences at any time point along the distribution, although sample size was not large (48 to 144 per condition). Since we assumed that a substantial part of the effect would show itself in the minimal response times, we needed a test that was especially sensitive to the extrema of a distribution. We finally settled on the AD test as it fulfilled these criteria better than the KS test.

Anderson-Darling Test

The Anderson-Darling test was developed in 1952 by T.W. Anderson and D.A. Darling (Anderson & Darling, 1952) as an alternative to other statistical tests for detecting sample distributions' departure from normality. Just like the KS test, it was originally intended and used mainly for engineering purposes.

The one-sample AD test statistic is non-directional, and is calculated from the following formula:

$$AD = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)(\ln(x_{(i)}) + \ln(1-(x_{(n+1-i)}))) \qquad (3)$$

where $\{x_{(1)} < \ldots < x_{(n)}\}$ is the ordered (from smallest to largest element) sample of size *n*, and *F(x)* is the underlying theoretical cumulative distribution to which the sample is compared. The null-hypothesis that $\{x_{(1)} < \ldots < x_{(n)}\}$ comes from the underlying distribution *F(x)* is rejected if AD is larger than the critical value AD_α at a given α (for a table of critical values for different sample sizes, see D'Agostino & Stephens, 1986).

The two-sample AD test, introduced by Darling (1957) and Pettitt (1976), generalizes to the following formula:

$$AD = \frac{1}{mn}\sum_{i=1}^{n+m}(N_i Z_{(n+m-ni)})^2 \frac{1}{iZ_{(n+m-i)}} \qquad (4)$$

where $Z_{(n+m)}$ represents the combined and ordered samples $X_{(n)}$ and $Y_{(m)}$, of size *n* and *m* respectively, and N_i represents the number of observations in $X_{(n)}$ that are equal to or smaller than the ith observation in $Z_{(n+m)}$. See Pettitt (1976) for critical values depending on α and sample size. The null hypothesis that samples $X_{(n)}$ and $Y_{(m)}$ come from the same continuous distribution is rejected if AD is larger than the correspondent critical value.

The AD test has been further generalized to a k-sample version (Scholz & Stephens, 1987), which is especially useful to test for the homogeneity of several samples. However, this version will not be discussed in this article.

Several comparisons between the one-sample AD test and other similar tests have been made. Anderson and Darling (1954) found that for one set of observations, the KS and

AD test produced the same result. Stephens (1974) compared several one-sample goodness of fit tests, and concluded that while all tests surpassed the χ^2 test in power, the KS, AD, and Cramer-von Mises tests detected changes in mean better.

The AD test has the same advantages mentioned for the KS test in the previous section, namely its sensibility to shape and scale of a distribution (Anderson & Darling, 1954) and its applicability to small samples (Pettitt, 1976). Specifically, the critical values for the AD test rise asymptotically and converge very quickly towards the asymptote (Anderson & Darling, 1954; Pettitt, 1976; Stephens, 1974).

In addition, the AD test has two extra advantages over the KS test. First, it is especially sensitive towards differences at the tails of distributions (as we will show next). Second, there is evidence that the AD test is better capable of detecting very small differences, even between large sample sizes. This is one of its main advantages in the field of engineering. The goal of the following Monte Carlo simulations is to investigate more rigorously the differences in performance between the KS test and the AD test, especially concerning small differences between samples and sensitivity to tail differences.

The AD test can be used in the same experimental context as the KS test, but it is not known in the field of psychology, the two-sample version even less than the one-sample version. Rare examples of use of the one-sample AD test in psychology include a test of normality for the distribution of judgments of verticality (Keshner, Dokka & Kenyon, 2006), and a test of normality of platelet serotonin level distributions (Mulder et al., 2004). Apart from our own studies (Engmann & Cousineau, submitted), we are not aware of any further examples of use of the two-sample version.

Comparison of the two tests when shift, scale and symmetry are varied independently

To compare the performance of KS versus AD test, we propose to test if the difference between two sets of data sampled from two minimally different distributions is statistically significant, according to the KS test and according to the AD test. By using theoretical distributions with known parameters, we are able to control the actual size of the difference between the two distributions. This allows us to compare the performance of both tests when distributions are very similar as well as when they are dissimilar. Also, this gives us a tool to observe the effect of change in specific parameters on the performance of both tests. Specifically, we can compare performance when distributions differ only at the extreme ends, but not around the mode, as will be done in the subsequent section.

Method

In a given simulation, we used two populations following Weibull distributions with three parameters,

$$D_1\ (\alpha, \beta, \gamma) \qquad (5a)$$

$$D_2\ (\alpha + \Delta_1, \beta + \Delta_2, \gamma + \Delta_3), \qquad (5b)$$

where $\alpha = 200$, $\beta = 80$, and $\gamma = 2.0$. These parameters are typical of speeded response time distributions (Heathcote, Brown & Cousineau, 2004). Δ_1 varied between –60 and 60, in steps of 4, Δ_2 varied between –30 and 30, in steps of 2, and Δ_3 varied between –1.2 and 1.2, in steps of 0.08. In the first simulations, only one parameter varied, whereas the other two remained the same ($\Delta = 0$). For each value of Δ_1, while maintaining Δ_2 and Δ_3 at 0, a

sample was drawn from D_1 as well as from D_2. A test of significant difference (with $\alpha = 0.05$) between D_1 and D_2 was then performed, using the KS test and then the AD test. This was repeated 10,000 times for each value of Δ_1 and subsequently for each value of Δ_2 and Δ_3 as well. For each value of Δ_1, Δ_2 and Δ_3 we were then able to calculate the probability of finding a significant difference between D_1 and D_2 for the KS test and for the AD test. This procedure was used for sample sizes of 16, 32 and 64, typical in experimental psychology.

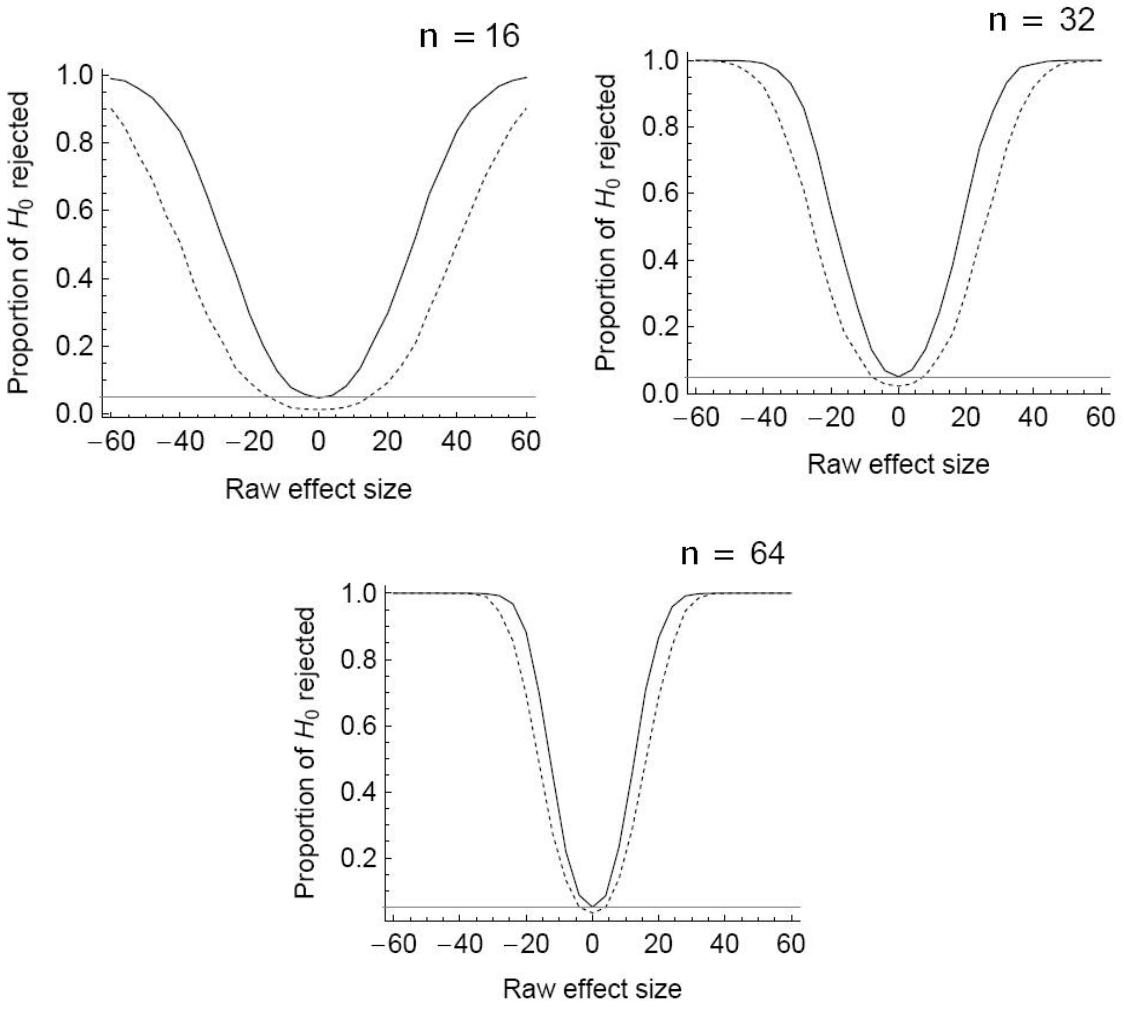

Figure 1. The proportion of significant differences between the two distributions for the AD and KS test as a function of Δ_1 (changes in shift). *The horizontal gray line is the boundary of an acceptable type I error rate for a decision criterion of 5%. Panels represent sample sizes 16, 32 and 64 respectively.*

Results

Figure 1 shows the probability for both AD test and KS test of finding a significant difference between D_1 and D_2 when Δ_1 changes, plotted along the abscissa. The three panels represent the different sample sizes. The probability of finding a significant difference is plotted as a function of Δ_1. If D_1 and D_2 are equal ($\Delta_1 = 0$), the AD test finds a significant difference (type I error) in 1.2% of the cases for sample size $n = 16$, 2.2% for $n = 32$, and

3.3% for n = 64. This is approximately the type I error usually allowed for (α). The KS test finds a significant difference in only 4.7% (n = 16), 5.0% (n = 32), and 5.2% (n = 64) of the cases. Hence, the KS test is slightly more conservative, allowing for a smaller proportion of type I errors. On the other hand, when Δ_1 differs from zero, the proportion of type II errors is larger for the KS test, finding no significant difference when distributions are actually different.

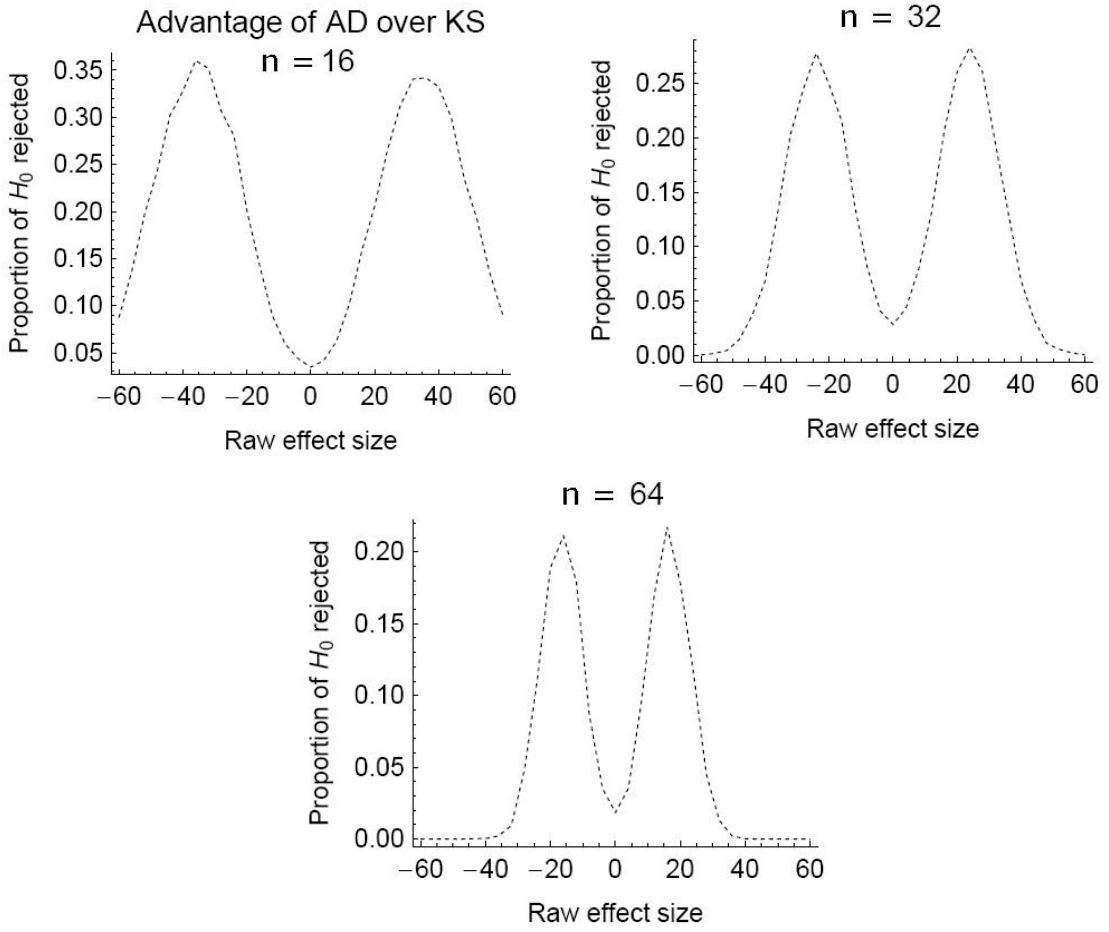

Figure 2. Absolute advantage of AD over KS test as a function of Δ_1 (changes in shift). *Panels represent sample sizes 16, 32 and 64 respectively.*

To illustrate the amount of gain of the AD test over the KS test more clearly, we calculated the difference in probability between the two tests. This was done by subtracting the KS-probability from the AD-probability of finding a significant difference for each value of Δ_1. Figure 2 plots the difference as a function of change in Δ_1, the panels representing sample sizes 16, 32 and 64 respectively. Figure 2 clearly shows that performance of the KS test approaches the performance of the AD test (i.e. the difference approaches zero) only for very large differences between distributions, or when the two distributions are equal (i.e. when Δ_1 = 0). As values of Δ_1 approach intermediate values (near ± 25), there is a systematic and constant gain, sometimes as large as 36% for the AD test over the KS test.

The AD test detects as much as a quarter of all differences for certain effect sizes which the KS test could not detect.

Differences in performance between KS test and AD test are more pronounced for small sample sizes. This holds for changes in Δ_1 as well as in Δ_2 and Δ_3, as will be shown next.

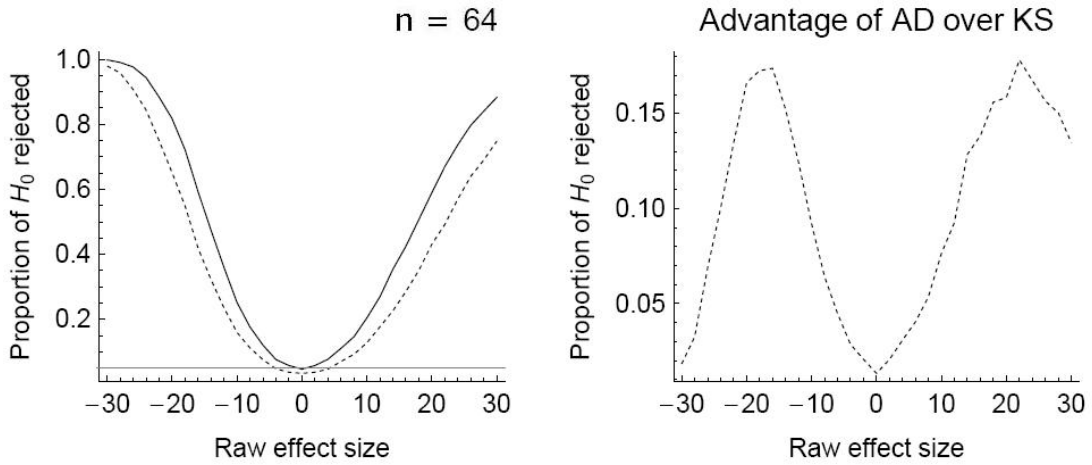

Figure 3. The proportion of significant differences between the two distributions for the AD and KS test as a function of Δ_2 (changes in scale). *The horizontal gray line is the boundary of an acceptable type I error rate for a decision criterion of 5%. The second panel shows the absolute advantage of the AD over the KS test.*

Figure 3a shows the probability for both AD test and KS test of finding a significant difference between D_1 and D_2 when Δ_2 changes, at a sample size of 64. When D_1 and D_2 were equal ($\Delta_2 = 0$), the proportion of type I errors for the AD test was 0.9% ($n = 16$), 2.0% ($n = 32$), and 3.3% ($n = 64$) respectively. Figure 3b represents the advantage of the AD test over the KS test, again at a sample size of 64. For all sample sizes, the AD test performed as good as or better than the KS test, with a maximal advantage of 4.2% ($n = 16$), 4.9% ($n = 32$) or 4.7% ($n = 64$) respectively.

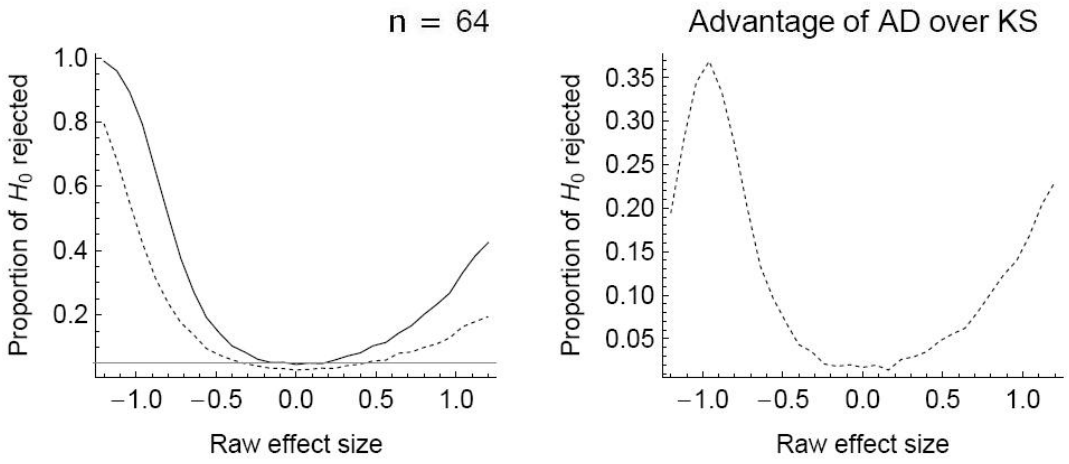

Figure 4: The proportion of significant differences between the two distributions for the AD and KS test as a function of Δ_3 (changes in asymmetry). *The horizontal gray line is the boundary of an acceptable type I error rate for a decision criterion of 5%. The second panel shows the absolute advantage of the AD over the KS test*

Figure 4a shows the probability for both AD test and KS test of finding a significant difference between D_1 and D_2 when Δ_3 changes, at a sample size of 64. The curve is less symmetrical as Δ_3 represents a change in symmetry, and the effect of a negative Δ_3 is not the same as the effect of a positive Δ_3. When D_1 and D_2 were equal ($\Delta_3 = 0$), the proportion of type I errors for the AD test was 0.7% ($n = 16$), 1.8% ($n = 32$), and 2.9% ($n = 64$) respectively. Figure 4b represents the advantage of the AD test over the KS test, again at a sample size of 64. For all sample sizes, the AD test performed as good as or better than the KS test, with a maximal advantage of 4.6% ($n = 16$), 4.9% ($n = 32$) or 4.6% ($n = 64$) respectively.

When D_1 and D_2 are equal, the KS test has a slightly lower type I error rate, but as soon as samples differ even slightly, the AD test outperforms the KS test for the detection of differences in shift (Δ_1), scale (Δ_2), or symmetry (Δ_3).

Comparison of the two tests when D_1 and D_2 differ in the tails only

As mentioned earlier, one of the strengths of the AD test is its sensitivity to the extreme ends of distributions – the minima and maxima. In order to test its performance specifically at the extrema, we decided to compare distributions that differed only at the extreme ends. The degree of difference between such distributions is extremely difficult to compute, and much less to control. Therefore we selected six instances of two distributions that differ at the extrema, and compared each with a KS and an AD test. One of these distributions was a Weibull, the other a Normal with approximately the same mean and variance as the Weibull. See Table 1 for the exact parameters of each of the six sets of distributions used. Figure 5 shows two such pairs of distributions. Weibulls can be asymmetrical, whereas Normals are symmetrical, which means that an overlap can be obtained for large parts of the distributions, while maintaining a difference at one or both of the extrema.

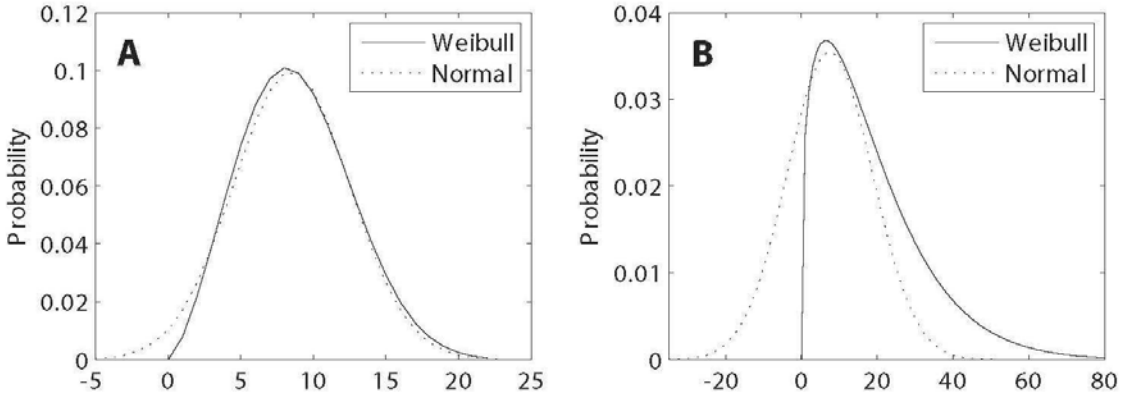

Figure 5. Weibull and Normal distributions used for evaluation of performance when distributions differ at tails. *The full line represents the Weibull, the dotted line the corresponding Normal distribution. Panel A shows the pair of distributions for which it was least likely to detect a difference (parameters: Weibull $\alpha = 0$, $\beta = 10$, $\gamma = 2.5$; Normal $\mu = 8.5$, $\sigma = 4$), panel B the pair for which it was most likely (parameters: Weibull $\alpha = 0$, $\beta = 20$, $\gamma = 1.3$; Normal $\mu = 7.5$, $\sigma = 11.5$).*

Method

We selected a sample of size 16 from the Weibull and the Normal in each set, tested them for significant difference using the KS and then the AD test. We repeated this procedure 10 000 times, and then calculated the probability of the AD and the KS test of finding a significant difference. In all other aspects, the procedure is the same as in the previous section.

Results

The results are shown in Table 1, the last column representing the gain of the AD over the KS test. The AD test is able to detect differences in distributions better than the KS test, even if they are located only at the tail(s) of a distribution.

Table 1. Parameters of the Weibull and Normal distributions from which samples are drawn for comparison. The last three columns show the probability (over 10 000 instances) of finding a significant difference between samples, either by the AD test or the KS test. The last column represents the advantage of the AD over the KS test.

	Weibull parameters			Normal parameters		Probability of finding a significant difference		
	α	β	γ	μ	σ	AD test	KS test	AD - KS
1	0	10	1.5	6	5.4	.200	.032	.168
2	0	10	2.5	8.5	4	.051	.005	.046
3	0	20	1.3	7.5	11.25	.561	.165	.396
4	0	20	4.0	17.5	6.75	.072	.013	.059
5	0	30	1.6	17.5	15.73	.252	.054	.198
6	0	30	2	22.5	14	.087	.018	.069

Sample size needed to reach sufficient statistical power when shift, scale and symmetry are varied independently

Another method to assess the advantage of one statistical method over another is based on statistical power (Cohen, 1992). We will compare the required number of data per cell to reach a target power. Following Cohen (1992), we will use 80% as the target power. The method which requires less data to reach a statistical power of 80% is to be preferred.

We defined the effect size for a shift relative to the standard deviation of the parent distribution. In the following, a small effect size is defined as a change in the shift (α) of the second distribution by a quantity of 0.25 σ and by a quantity of 0.75 σ for a large effect size. Table 2 lists the definitions of effect size for the three parameters. Hence, for a Weibull distribution with parameters γ = 2.0 and β = 80, the standard deviation is 37 ms and the small effect size is a shift by 9.3 ms (α ± 9.3 ms).

Table 2. Definition of large, medium and small effect size for the three parameters of the Weibull distribution

	Definition		
	Large	Medium	Small
α	0.75 σ	0.5 σ	0.25 σ
β	0.75 σ	0.5 σ	0.25 σ
γ	± 0.75	± 0.50	± 0.25

Regarding the scale parameter, there is no convention as to what constitutes a small, medium or large effect size. Hence, we adopted the same effect sizes for changes in scale as for changes in shift. Finally, for the changes in symmetry, a large effect size was defined as a change in the symmetry that would be clearly visible on a plot of the two distributions and a small effect as a change in the symmetry that would be difficult to see. As we saw in the first simulations, power is not symmetrical when the parameter γ is near 2.0. We chose to compare distributions with symmetry parameters of 1.25 and 2.75 (γ ± 0.75) for a large difference, 1.50 and 2.50 (γ ± 0.5) for a medium difference and finally 1.75 and 2.25 (γ ± 0.25) for a small difference. Figure 6 shows the resulting distributions for the two extreme conditions.

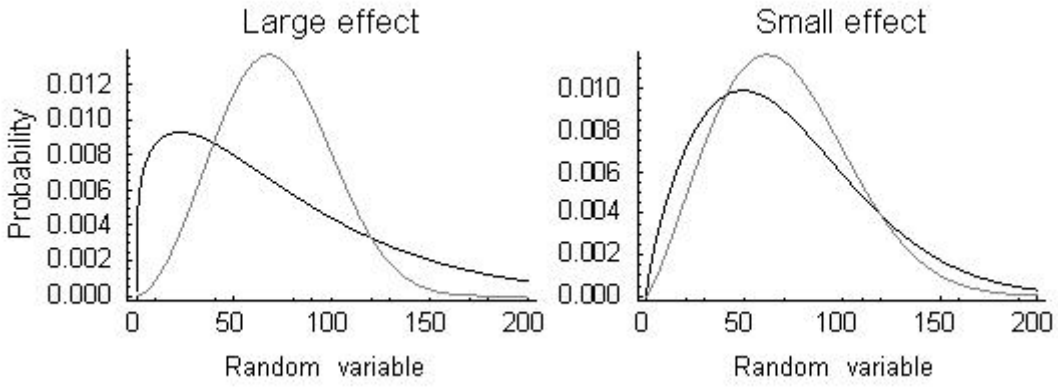

Figure 6. The two distributions compared when the effect size of change in symmetry is large (left) and small (right)

Method

Simulations were run in a fashion similar to the previous ones. We varied the sample size until a power of 80% was reached for each of the two tests, the AD test and the KS test. For most cases, the results are based on 10,000 simulations except when sample size is larger than 100, where the results are based on 25000 simulations so that the results are accurate to the third digit.

Results

The results are presented in Table 3. When the change is in the shift parameter, the net effect is to change the mean of the distribution. Hence, a powerful test should have about the same power as a standard test of means on two groups (e.g. a two-sample t-test). As seen, the number of data needed when the AD test is used (29, 61 and 233 for large, medium and small effect sizes respectively) is the same or slightly smaller than the number of data required by a t-test (29, 64 and 252 for large, medium and small differences in means; Cohen, 1992, Cousineau, 2007). The AD test is more powerful than a t-test when comparing two Weibull distributions; this can be explained by the fact that the left tail of a Weibull distribution is characterized by an abrupt onset. For a small effect size, there is an area of 9.3 ms where there are data in the first sample but none in the second sample. Since the AD test is sensible to differences in tails, it detects this difference in the left tail efficiently. When the two populations are normal, there is no advantage of the AD test over the t-test. The number of required data is 31, 69 and 272 for large, medium and small effect sizes respectively (based on Monte Carlo simulations with normal distributions).

Table 3 also shows the required number of data when the scale parameter and the symmetry parameter are varied. For changes in shift and scale, the required sample size by a KS test to obtain a statistical power of 80% is close to 50% larger than the sample size when using an AD test. Worst, the KS test is poorest at detecting changes in asymmetry, requiring almost twice as many data than the AD test.

Table 3. Number of data required to reach a power of 80% as a function of the effect size and the test used

	The Anderson-Darling test			The Kolmogorov-Smirnoff test		
	Large	**Medium**	**Small**	**Large**	**Medium**	**Small**
α	29	61	233	42	92	360
β	58	116	412	81	161	564
γ	48	100	377	83	190	768

In a regular psychology experiment, it is not known whether two groups differ with respect to their shape, scale, or symmetry, or a combination of the above. Hence, the following could be a reasonable rule of thumb for deciding the sample size to ensure sufficient statistical power: For a given expected effect size, choose the sample size associated with the parameter that requires the largest number of data. For example, if a medium difference is expected between two conditions, not knowing which parameter(s) will reflect the change, a safe approach would be to have 116 data per condition (a change in the scale parameter requires the highest number of data to ensure sufficient power). However, this ideal rule of thumb is limited by practical considerations: Considering that an experimental session generally has no more than 600 trials, that there may be a few erroneous responses that must be removed from the samples, and that there usually are more than two or three different conditions in an experiment, a sample size of 116 per condition might not be practical. If a KS test is used, this number reaches 190, a figure nearly impossible to obtain in any practical experimental design. Note that pooling data between sessions to increase sample size per condition is not recommended unless there are no significant practice effects.

Discussion

In conclusion, we have shown that the AD test is more powerful than the KS test in detecting any kind of difference between samples from two different distributions, all the while maintaining an exact type I error rate of .05. The KS test is overly conservative in comparison. This paper provides three different types of evidence that the performance of the AD test is superior. First, the AD test detects small variations of any one parameter between two distributions more reliably than the KS test. This holds for shift, scale and symmetry parameters, and for all sample sizes. Second, the AD test detects differences at the extreme ends of distributions more reliably than the KS test. Again, this holds even for small sample sizes and when the two distributions largely overlap. Finally, the AD test requires much less data per condition than the KS test in order to obtain sufficient statistical power. Since the AD test further possesses the same advantages as the KS test, and can be applied in the same experimental context, the evidence of its superior performance presented here shows that it should be preferred to the KS test as a tool for comparing distributions.

The AD test is recommended in any experimental context which requires a comparison of samples of continuous distributions, such as response time data, which requires more than a comparison of sample means.

The MatLab (MathWorks, Inc., Natick, MA) version of the two-sample Anderson-Darling test, "adtest2.m" for sample sizes larger than eight for both samples is provided in the Appendix. It requires as input two separate arrays of data, which do not need to be the same length. Samples are not required to be ordered before serving as input. Optionally, the type I error rate (α) can also be given as the third input. If omitted, the default value is α = .05. The output of "adtest2.m" confirms or rejects the null hypothesis that both samples come from the same underlying distribution, supplying the value of the AD statistic and the critical value for the specified α. Please note that the AD test is non-directional, that is it will only give evidence of a significant difference between samples, but not which one of the two is greater or smaller. For details on how to use the one-sample AD test, please refer to Stephens (1974).

Acknowledgments

The authors would like to thank Sophie Callies, Étienne Dusmesnil and Laurence Morissette for their comments on a previous version of the manuscript. This research was supported by the *Conseil pour la recherche en sciences naturelles et en génie du Canada*, the *German Academic Exchange Service (DAAD)* and the *Friedrich-Ebert-Stiftung*.

References

1. Anderson, T. W., Darling, D. A. **Asymptotic theory of certain goodness-of-fit criteria based on stochastic processes**. Annals of Mathematical Statistics, 23, 193-212, 1952
2. Anderson, T. W., Darling, D. A. **A test of goodness of fit.** Journal of the American Statistical Association, 49, 765–769, 1954
3. Chernoff, H., Lehmann, E. L. **The use of maximum likelihood estimates in χ^2 tests for goodness-of-fit.** The Annals of Mathematical Statistics, 25, 579-586, 1954
4. Cohen, J. **A power primer.** Psychological Bulletin, 112, 155-159, 1992
5. Cousineau, D. **Implementing and evaluating the nested maximum likelihood estimation technique.** Tutorials in Quantitative Methods for Psychology, 3, 8-13, 2007
6. Cousineau, D., Brown, S., Heathcote, A. **Fitting distributions using maximum likelihood: Methods and packages**, Behavior Research Methods, Instruments, & Computers, 36, 742-756, 2004
7. D'Agostino, R. B., Stephens, M. A. **Goodness-of-Fit Techniques.** New York: Marcel Dekker, 1986
8. Darling, D. A. **The Kolmogorov-Smirnov, Cramér-von Mises tests**. The Annals of Mathematical Statistics, 28, 823-838, 1957
9. Eckel, C. C., Grossman, P. J. **Are women less selfish than men? Evidence from dictator experiments.** Economic Journal, 108, 726–735, 1998
10. Engmann, S., Cousineau, D. (submitted). **Coactivation results cannot be explained by pure coactivation models**.
11. Galambos, J. **The Asymptotic Theory of Extreme Order Statistics**, New York: John Wiley and Sons, 1978
12. Gumbel, E. J. **The Statistics of Extremes**, New York: Columbia University Press, 1958
13. Heathcote, A., Brown, S., Cousineau, D. **QMPE: Estimating Lognormal, Wald and Weibull RT distributions with a parameter dependent lower bound**. Behavior Research Methods, Instruments, & Computers, 36, 277-290, 2004

14. Isaic-Maniu, Al. **Metoda Weibull - aplicatii**, Ed. Academiei, Bucharest, 1983
15. Keshner, E. A., Dokka, M. S., Kenyon, R. V. **Influences of the Perception of Self- Motion on Postural Parameters.** Cyberpsychology and Behaviour, 9(2), 163-166, 2006
16. Kolmogorov, A. N. **Sulla determinazione empirica di une legge di distribuzione.** Giornale dell'Istituto Italiano degli Attuari, 4, 83-91, 1933
17. Kolmogorov, A. N. **Confidence limits for an unknown distribution function.** Annals of Mathematical Statistics, 12, 461-463, 1941
18. Lacouture, Y., Cousineau, D. (in press). **How to use MATLAB to fit the ex-Gaussian and other probability functions to a distribution of response times.** Tutorials in Quantitative Methods for Psychology
19. Lilliefors, H. W. **On the Kolmogorov-Smirnov test for normality with mean and variance unknown.** Journal of the American Statistical Association, 62, 399-402, 1967
20. Lilliefors, H. W. **On the Kolmogorov-Smirnov test for the exponential distribution with mean unknown.** Journal of the American Statistical Association, 6 , 387-389, 1969
21. Logan, G. D. **Shapes of reaction-time distributions and shapes of learning curves: a test of the instance theory of automaticity**. Journal of Experimental Psychology: Learning, Memory and Cognition, 18, 883-914, 1992
22. Massey, F. J. **The Kolmogorov-Smirnov test of goodness of fit.** Journal of the American Statistical Association, 46, 68-78, 1951
23. Miller, J. **Divided attention: Evidence for coactivation with redundant signals.** Cognitive Psychology, 14, 247-279, 1982
24. Mordkoff, J. T., Yantis, S. **An interactive race model of divided attention.** Journal of Experimental Psychology: Human Perception and Performance, 17, 520-538, 1991
25. Mordkoff, J. T., Yantis, S. **Dividing attention between color and shape: Evidence of co-activation.** Perception and Psychophysics, 53, 357-366, 1993
26. Mulder, E. J., Anderson, G. M., Kema, I. P., de Bildt, A., van Lang, N. D. J., den Boer, J. A., Minderaa, R. B. **Platelet Serotonin Levels in Pervasive Developmental Disorders and Mental Retardation: Diagnostic Group Differences, Within-Group Distribution, and Behavioral Correlates**. Journal of the American Academy of Child and Adolescent Psychiatry, 43(4), 491-499, 2004
27. Pandit, A. **Circadian rhythm variation in attempted suicide by deliberate self-poisoning, and in completed suicide, in central Nepal.** Biological Rhythm Research, 35(3), 229-233, 2004
28. Pettitt, A. N. **A two-sample Anderson-Darling rank Statistic.** Biometrika, 63, 161-168, 1976
29. Rodriguez, R., Campos-Sepulveda, E., Vidrio, H., Contreras, E., Valenzuela, F. **Evaluating knowledge retention of third year medical students taught with an innovative pharmacology program.** Academic Medicine, 77(6), 574-577, 2002
30. Rouder, J. N., Lu, J., Speckman, P., Sun, D., Jiang, Y. **A hierarchical model for estimating response time distributions.** Psychonomic Bulletin & Review, 12, 195-223, 2005
31. Rouder, J. N., Speckman, P. L. **An evaluation of the Vincentizing method of forming group-level response time distributions.** Psychonomic Bulletin and Review, 11, 419-427, 2004
32. Scholz, F. W., Stephens, M. A. **K-sample Anderson-Darling Tests.** Journal of the American Statistical Association, 82(399), 918-924, 1987
33. Smirnoff, H. **Sur les Écarts de la Courbe de la Distribution Empirique.** Receuil Mathémathique (Matematiceskii Sbornik), 6, 3-26, 1939
34. Stephens, M. A. **EDF statistics for goodness of fit and some comparisons.** Journal of the American Statistical Association, 69, 730-737, 1974

35. Thomas, E. A. C., Ross, B. **On appropriate procedures for combining probability distributions within the same family.** Journal of Mathematical Psychology, 21, 136-152, 1980

36. Vincent, S. B. **The function of vibrissae in the behavior of the white rat.** Behavioral Monographs, 1(5), 1912

Appendix

Implementation of the two-sample Anderson-Darling test in MatLab (MathWorks, Inc., Natick, MA). This implementation assumes sample sizes to be larger than eight. Please refer to D'Agostino and Stephens (1986) for an approximate adjustment of the calculation of the AD statistic for smaller sample sizes, or to Pettitt (1976) for a table of critical values of the AD statistic for smaller sample sizes.

```
function [H, adstat, critvalue] = adtest2(sample1, sample2, alpha)
% ADTEST2: Two-sample Anderson-Darling test of significant difference.
% This test is implemented for sample sizes larger than 8. For smaller
% sample sizes please refer to A.N.Pettitt, 1976 (A two-sample
% Anderson-Darling rank statistic) for the critical values.
%
% CALL:                 adtest2 (sample1, sample2);
%      [H,adstat,critvalue] = adtest2 (sample1, sample2, alpha);
%      Sample1 and sample2 are the samples to be compared. They must
%      be vectors of a size greater than 8. Alpha specifies the
%      allowed error. If alpha is not specified, a default value of
%      0.05 for alpha is used. Alpha must be either 0.01, 0.05 or 0.1.
%
% RETURN: H gives the statistical decision. H = 0: samples are not
%         significantly different. H = 1: sample1 and sample2 are
%         significantly different (i.e. do not arise from the same
%         underlying distribution).
%       adstat returns the ADstatistic of the comparison of the two
%         samples. If adstat is greater than the critical value,
%         the two samples are significantly different.
%       critvalue returns the critical value for the alpha used
%
% (c) Sonja Engmann 2007

if nargin < 2, error('Call adtest2 with at least two input arguments'); end
if nargin < 3, alpha = 0.05; end

% Assignment of critical value depending on alpha
if alpha == 0.01, critvalue = 3.857;
elseif alpha == 0.05, critvalue = 2.492;
elseif alpha == 0.1, critvalue = 1.933;
else error('Alpha must be either 0.01, 0.05 or 0.1.');
end

samplecomb = sort([sample1 sample2]);
ad = 0;
for i = 1:length(samplecomb)-1
   m = length(find(sample1(:)<=samplecomb(i)));
   ad = ad + (((m*length(samplecomb) - length(sample1)*i)^2)/(i*(length(samplecomb)-i)));
end
adstat = ad/(length(sample1)*length(sample2));
if adstat > critvalue, H = 1; else H = 0; end
```

[1]Sonja ENGMANN holds a B. Sc. in experimental psychology from University of Osnabrück, Osnabrück, Germany and a Ph. D. in cognitive psychology from Université de Montréal.

[2] Denis COUSINEAU is professor at University of Ottawa, Canada. He holds a B. Sc. in computer science, and another one in psychology. His Ph. D. in cognitive psychology was obtained at Université de Montréal in 1999 and he completed a post-doc at Indiana University. He specializes in mathematical psychology, in neural networks, and studies the human object recognition mental processes.

JOURNAL OF APPLIED QUANTITATIVE METHODS

Quantitative Methods Inquires

THE CREDIBILITY, COMPLETENESS AND ACCURACY OF INFORMATION ABOUT FIRST AID IN CASE OF CHOKING ON THE ROMANIAN WEBSITES

Valentin NADASAN
Department of Hygiene and Community Nutrition,
University of Medicine and Pharmacy, Tirgu Mures, Romania

E-mail: vnadasan@gmail.com

Gheorghe VANCEA
Department of Anesthesiology and Intensive Care,
Mures County Emergency Hospital, Tirgu Mures, Romania

Paula-Alina GEORGESCU
Department of Hygiene and Community Nutrition,
University of Medicine and Pharmacy, Tirgu Mures, Romania

Monica TARCEA
Department of Hygiene and Community Nutrition,
University of Medicine and Pharmacy, Tirgu Mures, Romania

Zoltan ABRAM
Department of Hygiene and Community Nutrition,
University of Medicine and Pharmacy, Tirgu Mures, Romania

Abstract: *A large number of studies assessing the quality of medical websites in various languages have shown that the quality of health related information is problematic. Nevertheless, the Romanian medical cyberspace has not yet been systematically evaluated. The goal of our study was to assess the credibility and content quality of information about first aid in case of choking intended for the general population on the Romanian websites. We evaluated a sample of 20 websites selected from the Google's first search results pages. The compliance to the credibility criteria was very low, the coverage of the topic was medium and the accuracy was good but we found frequent omission of important information about the first aid procedure for choking. Websites with high completeness and accuracy scores were rare, therefore, users should check several websites in order to get thoroughly and correctly informed about the topic.*

Key words: *First aid, choking, Heimlich maneuver, information quality, Internet*

Introduction

The use of Internet as a source of health related information has grown continually in the last decade [1]. A large number of studies assessing the quality of English, Spanish and French medical websites have shown that a considerable proportion of them are of poor quality and users are exposed to significant risks by taking wrong decisions about their treatment procedures [2]. Nevertheless, the Romanian medical cyberspace has not yet been systematically evaluated, except for a preliminary study dealing with the general characteristics of the medical websites [3].

The purpose of the present research was to assess the quality of information about first aid instruction for choking on the Romanian websites intended for nonprofessionals. We tried to answer the following questions: (a) What is the degree in which the websites presenting the first aid in case of choking comply with the European credibility criteria? (b) How completely and accurately is the topic covered? (c) Are there any website characteristics associated with poor/high quality information? (d) Is the level of compliance to the European credibility criteria correlated to the quality of the websites' content?

.

Data and methods

We included in our sample the first 20 websites listed by Google on the first five results pages [4,5]. We did not use Google's advanced search features but we did limit the search to the Romanian webpages by initiating the search at URL: www.google.ro. We have used successively the following search therms: "*Manevra Heimlich*" ("Heimlich maneuver"), "*Prim ajutor sufocare*" ("First aid choking"), "*Dezobstrucţia căilor respiratorii*" (Airways desobstruction"). The search was done during June-August 2011. We included only those sites that covered the topic in at least 250 words in Romanian language and which targeted the general population. We excluded all sponsored links, discussion forums, infected or unavailable sites and also sites that required registration. If several pages or subdomains belonging to the same top level domain were listed as separate links on the search engine's results page, we examined them as one website.

We classified the websites by their general characteristics (type of ownership, main purpose, genre and medical paradigm) [6,7]. Then, the websites were screened for compliance to a set of 14 quality criteria derived from the eEurope 2002 quality principles [8]. The evaluation included 14 questions along with detailed instructions for the reviewers. (The form is available upon request from the first author). For every criterion that was met, the website was awarded one point. The sum of all points resulted in the eEurope credibility score (eS) of the respective website. Next, the content of each website was checked against a list of expected items that we developed from the first aid guidelines issued by local and international professional organizations. This standard content list was also included in an assessment form along with comprehensive instructions for the evaluators. (This form is also available upon request from the first author). For each standard item covered the website was granted one point, regardless of the accuracy of the information [9]. The total number of items addressed resulted in what we called the absolute completeness score (aCS) of the website. Each item addressed on the site was then rated for accuracy, on a three level scale: totally correct (2 points), mostly correct (1 point), mostly incorrect (0 points) [9,10]. The sum of all points awarded to a site resulted in the absolute accuracy score (aAS). In order to

enable comparison of the results with those of other studies on health topics having different number of items on the standard list, we calculated the relative completeness score (rCS) and, respectively, the relative accuracy score (rAs) as shown below:

rCS = 10 aCs / mCS (where, **mCS** represents the maximum completeness score (identical to the total number of items on the standard list).

Likewise:

rAS = 10 aAS/mAS (where, **mAS** represents the maximum number of points that a specific site could be awarded supposing all the items addressed were totally accurate (site specific maximum accuracy score). The values of both relative scores could thus vary from a minimum of 0 to a maximum of 10.

We also calculated a risk score (RS) that was measured by counting the total number of items that could pose a health risk for the users, either by omission or by commission [11, 12].

All websites were rated by two independent evaluators who followed the common set of instructions provided in the assessment form. The data were centralized, compared for discrepancies and all disagreements were settled by consensus.

We checked for statistical differences between the quality scores of the websites classified by their general characteristics with the nonparametric Mann-Whitney (U) test or Kruskal-Wallis test (.05 level of significance)[13], and also the correlation between the eEurope credibility score and content quality scores with Spearman rank correlation test [14]. All statistical analyses were carried out using Graphpad InStat Demo 3.06.

Results

Regarding the type of ownership, most of the websites (12 out 20) were owned by commercial companies, four were owned by a foundation, a private medical institution, a state medical institution and an individual, respectively, and four more had unidentified owners. As far as the main purpose most of the websites (18 out of 20) were educational and two commercial. As far as the genre of the site, the sample included: one blog, one forum, one topical website, three general webportals, three company presentation websites, five medical webportals, and six online magazines/journals. Finely, taking in consideration the medical paradigm, we identified three complementary and alternative medicine websites and seventeen conventional medicine websites.

The average eEurope credibility score (eS) was 4.90 (SD 1.71). The distribution of the credibility scores across the 20 websites is represented in figure 1. The percentage of sites complying to each individual eEurope 2002 credibility criteria is shown in figure 2.

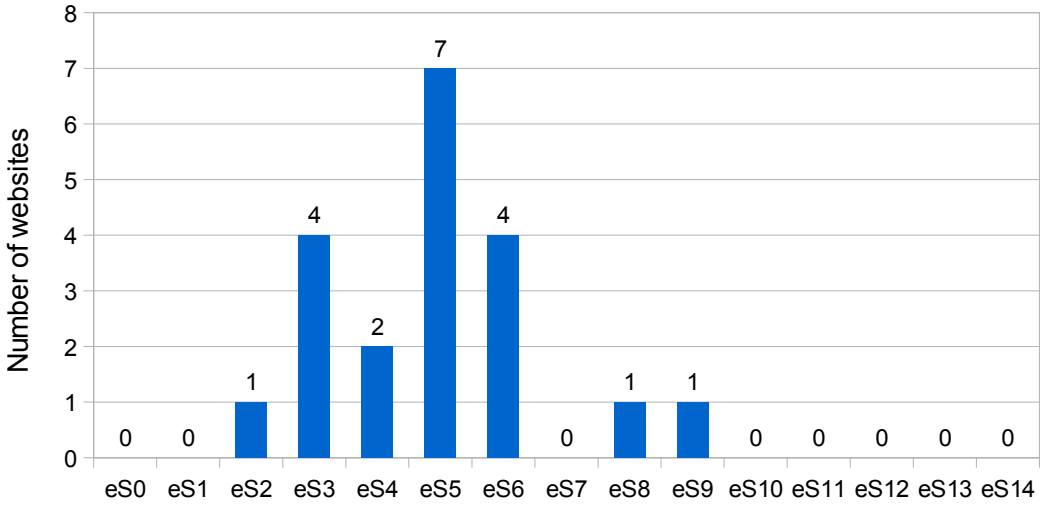

Figure 1. Distribution of eEurope 2002 credibility scores (eS)

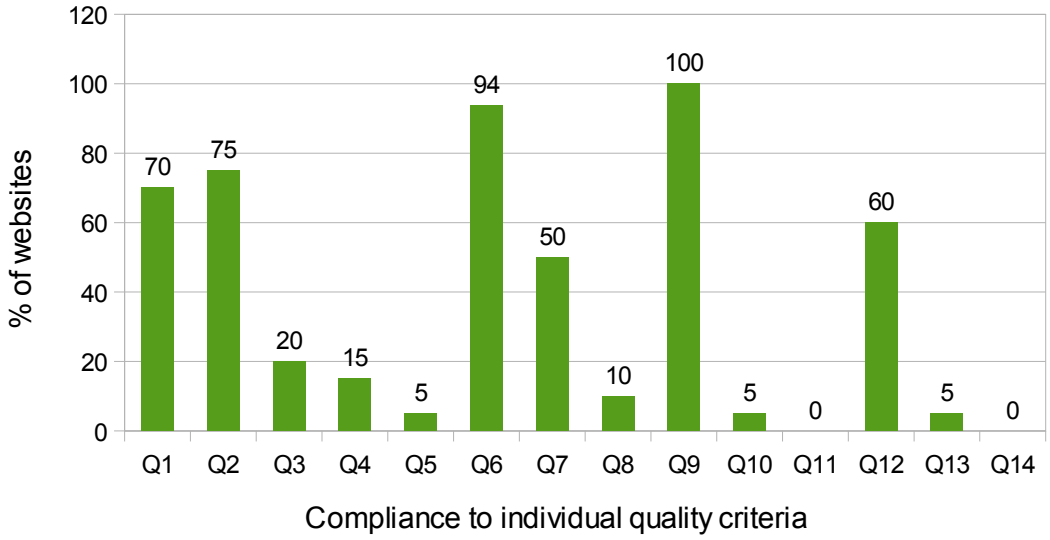

Figure 2. Compliance to individual eEurope 2002 quality criteria

The average relative completeness score (rCS) of the sample was 6.10 (SD 1.55) and the average accuracy score (rAS) was 8.56 (SD 0.99). The distribution of the rCS and rAS among the 20 websites evaluated for information on first aid in case of choking is represented in figure 3.

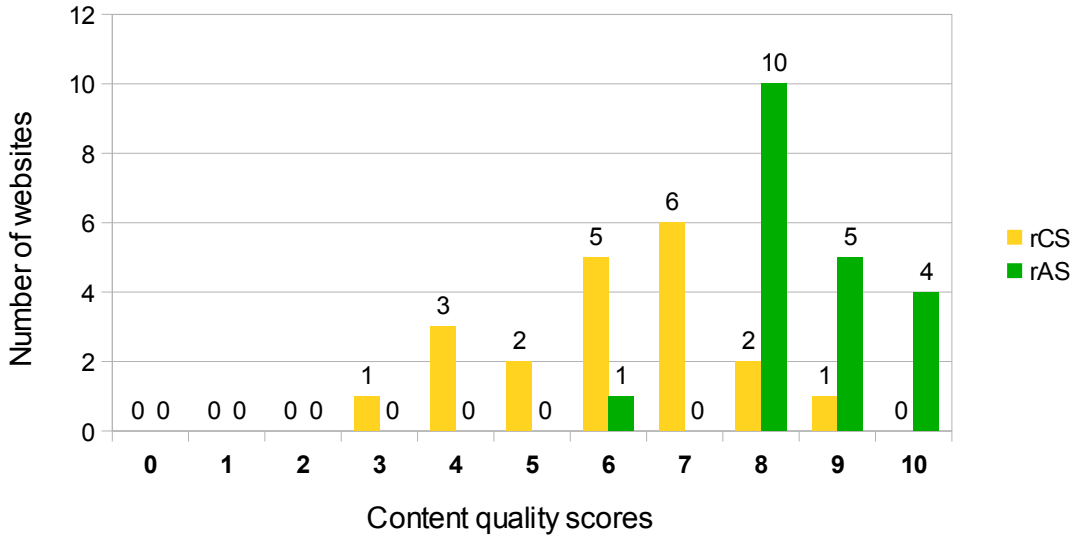

Figure 3. Distribution of relative completeness scores (rCS) and relative accuracy scores (rAC)

The average risk score (RS) of the websites was 4.75 (SD 2.00). The distribution of the RS among the examined websites is represented in figure 4.

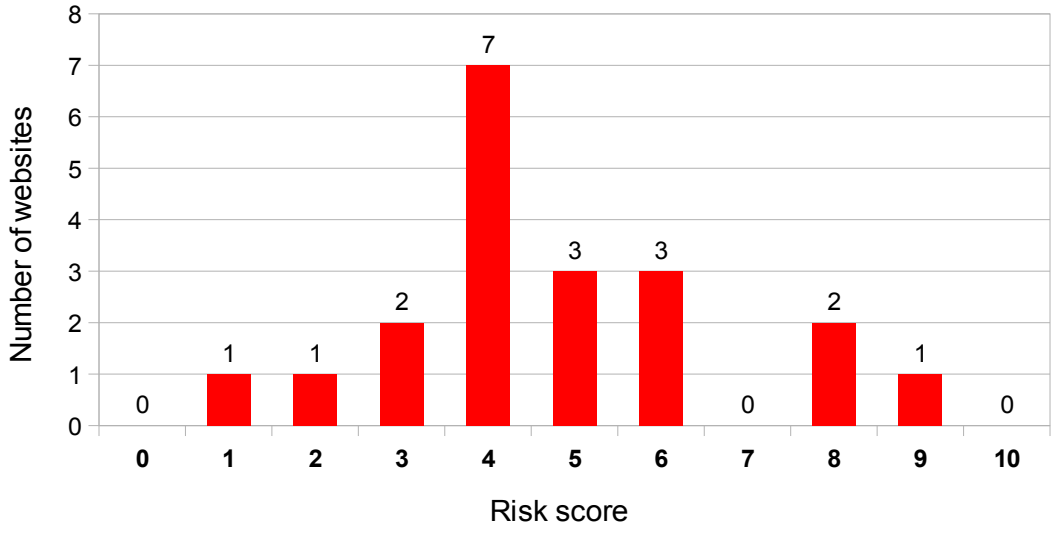

Figure 4. Distribution of the risk scores (RS)

The P values found after checking the statistical differences between the eEurope 2002 compliance scores, completeness, accuracy and risk scores of the websites classified according to their general characteristics, are shown in table 1.

Table 1. The P values for differences between the scores of the sites classified according to their general characteristics

Variables		Test	P value	Interpretation
Ownership*	eS	Kruskal-Wallis	0.0581	Not significant
	rCS	Kruskal-Wallis	0.3032	Not significant
	rAS	Kruskal-Wallis	0.6570	Not significant
	RS	Kruskal-Wallis	0.5756	Not significant
Purpose	eS	Mann-Whitney	0.4479	Not significant
	rCS	Mann-Whitney	0.7527	Not significant
	rAS	Mann-Whitney	0.6570	Not significant
	RS	Mann-Whitney	0.6570	Not significant
Genre*	eS	Kruskal-Wallis	0.0145	Significant
	rCS	Kruskal-Wallis	0.8901	Not significant
	rAS	Kruskal-Wallis	0.6158	Not significant
	RS	Kruskal-Wallis	0.3307	Not significant
Medical paradigm	eS	Mann-Whitney	0.5584	Not significant
	rCS	Mann-Whitney	0.3958	Not significant
	rAS	Mann-Whitney	0.6684	Not significant
	RS	Mann-Whitney	0.9999	Not significant

eS = eEurope credibility score rCS = relative completeness score
rAS = relative accuracy score RS = risk score
* Some of the original categories were merged to meet testing conditions.

The Spearman "r" statistics and the corresponding P values found after checking for correlations between the eEurope credibility scores and the content quality scores are presented in table 2.

Table 2. Spearman statistics for correlations between credibility and content quality scores

Variables		r*	P value	Interpretation
eS	rCS	0.3081	0.1863	Weak positive correlation; P value considered not significant.
eS	rAS	-0.2465	0.2947	Very weak or no correlation; P value considered not significant.
eS	RS	-0.1240	0.6026	Very weak or no correlation; P value considered not significant.

* r values were corrected for ties

Discussions

To the best of our knowledge this is the first study assessing the quality of information about first aid instructions in case of choking on the Romanian websites intended for non-professionals.

The average credibility score (4.90 out of 14) suggests that the overall compliance of the Romanian websites addressing the topic under evaluation to the European quality criteria is very low. Of all the website categories, we found that only medical portals seem to have a significantly higher compliance score compared to the other genre of websites with health related information. The compliance to the eEurope credibility criteria varies greatly from one criterion to the other. Providing a feedback form and differentiating the advertising from editorial content have the highest rate of compliance (100% and 94% respectively). Other criteria with fairly good compliance are the statement of purpose (75%) and the disclosure of ownership (70%). At the opposite end, several important criteria have low or

very low compliance levels: disclosure of sponsorship (15%), financial interest (5%), authors name and credentials (5%), providing the date of the last update (5%), references (0%), and editorial review policy (0%). Compared to the data reported by Eysenbach et al. in one of the most comprehensive systematic reviews, the Romanian medical websites seem to have a higher level of compliance to the following credibility criteria: statement of purpose (75% vs. 48%), disclosure of sponsorship (15% vs. 6%), differentiation of advertisement from editorial content (94% vs. 69%), providing a feedback mechanism (100% vs 86%), and disclosure of the first publication date (60% vs. 17%). On the other hand, the health related websites in our sample seem to have lower level of compliance to some other credibility criteria such as: disclosure of authorship and their credentials (5% vs. 30%), providing references (0% vs. 31%) and disclosure of the editorial policy (0% vs. 13%) [9].

The average relative completeness score (6.10 out of 10) indicates that the Romanian websites coverage of the investigated topic is, at best, medium. Although rigorous comparison of results would be difficult because of some methodological reasons, many of the published papers on the quality of medical information on the English and Spanish websites about a wide range of health topics such as scoliosis[15], cervical disc herniation[16], breast cancer, childhood asthma, depression, obesity[14], cocaine addiction[17], diabetes[18], nutrition[19], arrive to the conclusion that the coverage of these topics is problematic. Eysenbach et al. also note in their review of the literature that „most authors who evaluated content, found significant problems, criticizing lack of completeness" and specifically mention that five of eight studies reviewed, reported that around 90% of the websites were „incomplete" [9].

The accuracy of information about first aid instructions for choking on the Romanian Web seems good if judged by the average relative accuracy score of 8.56. In contrast, most of the published literature about the quality of information on various health or disease related topics in English language shows a low level of accuracy [14-19]. The percentage of inaccurate websites reported by Eysenbach et al in their review varies widely (4-9% among cancer websites up to 45-88% among diet and nutrition websites) [9]. However, the notable methodological heterogeneity of the studies permit only a very loose comparison.

It is important to note that the completeness and accuracy scores as applied in our study must not be interpreted independently because the completeness score was intended to measure exclusively the coverage of the topic and the accuracy score the correctness of information without any reference to completeness. As such, websites with extremely low coverage of the topic, can get high or very high accuracy scores if the information they present, as little as may be, is correct.

Only three of the twenty sites had both high completeness and accuracy scores (rCS and rAS > 8 points). Therefore the probability of finding exhaustive and simultaneously correct information about the topic is rather low unless the users are looking for information on more than one site.

The risk score indicates that omissions of important facts are frequent. Some of the websites fail to warn the users on as many as 7 items of critical importance for the victims of choking. The most notable deficiency is that all the websites except for one, are missing step number one of the first aid procedure in case of choking namely, the delivery of five back blows between the person's shoulder blades with the heel of the hand. Apparently, many websites describing the first aid procedure make extensive use of outdated information promoted by the Heimlich Institute.

None of the correlations between the eEurope credibility score and the content quality scores (rCS, rAS, RS) reached statistical significance, therefore, our study suggests that the credibility criteria are not helpful for the general users in identifying scientifically accurate websites. Our results seem to be in line with the previous reports on this issue [13,20].

The main limitations of the study are those inherent to Internet research. First of all, the extreme dynamics of the cyberspace makes the exact replication of any study virtually impossible. Substituting the search terms could also significantly change the components and structure of the sample and thus the quality scores as well.

The relatively small number of websites included in our sample should not be necessarily regarded as a limitation, because more than one study has revealed that most typical Internet users don't look beyond the links on the first results page of the search engine anyway [4,21].

Although we tried to minimize the subjectivity of the assessment by providing to the evaluators all the elements that could possibly be anticipated, and also by carrying out the examination by two independent evaluators, we assume that results of our assessment suffer in a certain degree because of this factor. The score most likely to be dependent on the evaluators' subjectivity is the risk score and it has to do with those items that cannot be included a priori in the assessment form and are to be identified and judged by the evaluator based on their own medical knowledge.

Our study focused on an important but narrow domain of the Romanian medical cyberspace. In order to get a more comprehensive picture about the quality of health related information on the Romanian Internet, the spectrum of investigation should include the assessment of websites addressing a wide diversity of topics.

Conclusions

1. The level of compliance of the Romanian websites addressing the topic of first aid in case of choking to the credibility criteria was very low.

2. Overall, the coverage of the topic was medium while the accuracy of the addressed information was good. However, exhaustive and accurate websites about the investigated topic were rare and many websites were characterized by frequent omission of important information.

3. We found no statistically significant correlation between the credibility score and the content quality scores which makes unpractical the attempt to identify accurate websites based on their credibility features.

References

1. Eysenbach, G. **Consumer health informatics,** BMJ, 320, 7251, 2000, pp. 1713-6
2. Eng, T.R. and Gustafson, D.H. (editors), **Science Panel on Interactive Communication and Health. Wired for Health and Well-Being: the Emergence of Interactive Health Communication,** US Department of Health and Human Services, U.S. Government Printing Office, 1999, pp. 11/7-11/16.
3. Nadasan, V., Ancuceanu, R., Tarcea, M., Grosar, C.M., and Ureche, R. **General characteristics of the Romanian medical webscape,** Acta Medica Marisiensis, 57, 2, 2011, pp. 94-97

4. Schwartz, K.L., Roe, T., Northrup, J., Meza, J., Seifeldin, R. and Neale, A.V. **Family medicine patients' use of the Internet for health information: a MetroNet study,** J Am Board Fam Med. 19, 1, 2006, pp. 39-45
5. Institutul Român pentru Evaluare şi Strategie. **Românii şi internetul - martie 2011. Studiu privind utilizarea Internetului şi comportamentul internautic al românilor,** [Online] Retrieved: 05.15.2011. Available from: http://www.ires.com.ro/uploads/articole/ires_romanii_si_internetul_2011.pdf
6. Lawrentschuk, N., Abouassaly, R., Hackett, N., Groll, R. And Fleshner, N.E. **Health information quality on the internet in urological oncology: a multilingual longitudinal evaluation,** Urology, 74, 5, 2009, pp. 1058-63
7. Molassiotis, A. and Xu, M. **Quality and safety issues of web-based information about herbal medicines in the treatment of cancer,** Complement Ther Med, 12, 4, 2004, pp. 217-27
8. *** **eEurope 2002: Quality Criteria for Health related Websites,** [Online] J Med Internet Res, 4, 3, 2002,e15. Retrieved: 06.26.2011. Available from: http://www.jmir.org/2002/3/e15/
9. Eysenbach, G., Powell, J., Kuss, O. and Sa, E.R. **Empirical studies assessing the quality of health information for consumers on the world wide web: a systematic review,** JAMA, 287, 20, 2002, pp. 2691-700
10. Killeen, S., Hennessey, A., El Hassan, Y., Killeen, K., Clarke, N., Murray, K. et al. **Gastric cancer-related information on the Internet: incomplete, poorly accessible, and overly commercial,** Am J Surg, 201, 2, 2011, pp. 171-8
11. Walji, M., Sagaram, S., Sagaram, D., Meric-Bernstam, F., Johnson, C., Mirza, N.Q. et al. **Efficacy of quality criteria to identify potentially harmful information: a cross-sectional survey of complementary and alternative medicine web sites,** J Med Internet Res, 6, 2, 2004, e21
12. Markman, M. **Safety issues in using complementary and alternative medicine,** J Clin Oncol, 20, 18Suppl, 2002 pp. 39S-41S
13. Griffiths, K.M. and Christensen, H. **Quality of web based information on treatment of depression: cross sectional survey,** BMJ, 321, 7275, 2000, pp. 1511-5
14. Berland, G.K., Elliott, M.N., Morales, L.S., Algazy, J.I., Kravitz, R.L., Broder, M.S. et al. **Health information on the Internet: accessibility, quality, and readability in English and Spanish,** JAMA, 285, 20, 2001, pp. 2612-21
15. Mathur, S., Shanti, N., Brkaric, M., Sood, V., Kubeck, J., Paulino, C. et al. **Surfing for scoliosis: the quality of information available on the Internet,** Spine, 30, 23, 2005, pp. 2695-700
16. Morr, S., Shanti, N., Carrer, A., Kubeck, J. and Gerling, M.C. **Quality of information concerning cervical disc herniation on the Internet,** Spine, 10, 4, 2010, pp. 350-4
17. Khazaal, Y., Chatton, A., Cochand, S. and Zullino, D. **Quality of Web-based information on cocaine addiction,** Patient Educ Couns, 72, 2, 2008, pp. 336-41
18. Bedell, S.E., Agrawal, A. and Petersen, L.E. **A systematic critique of diabetes on the world wide web for patients and their physicians,** Int J Med Inform, 73, 9-10, 2004, pp. 687-94
19. Sutherland, L.A., Wildemuth, B., Campbell, M.K. and Haines, P.S. **Unraveling the web: an evaluation of the content quality, usability, and readability of nutrition web sites,** J Nutr Educ Behav, 37, 6, 2005, pp. 300-5
20. Kunst, H., Groot, D., Latthe, P.M., Latthe, M. and Khan, K.S. **Accuracy of information on apparently credible websites: survey of five common health topics,** BMJ, 324, 7337, 2002, pp. 581-2
21. Eysenbach, G. and Köhler, C. **How do consumers search for and appraise health information on the world wide web? Qualitative study using focus groups, usability tests, and in-depth interviews,** BMJ, 324, 7337, 2002, pp. 573-7

INDICATORS FOR ECONOMIC AND SOCIAL DEVELOPMENT OF FUTURE SMART CITY

Lorena BATAGAN[1]

PhD, Lecturer, Department of Economic Informatics,

University of Economics, Bucharest, Romania

E-mail: lorena.batagan@ie.ase.ro

Abstract: *This paper identifies a set of criteria and indicators to highlight the economic and social development from different smart cities. To characterize and prioritize various urban areas of a city should be used a set of indicators. These indicators help us to establish the evolution and the grade of prepare of cities to smart society. This represents a starting point in formulating objectives and strategies for smart city development.*

Key words: smart city; indicators; local development; regional development

1. Introduction

This paper has in view the medium-sized cities and their perspectives for development.

In our days most of the population lives in urban areas, so exploring the development and evolution of cities is essential. The challenge of medium-sized cities in an important urban area is an important step. Medium-sized cities, which in next year's will be near the larger metropolises, appear to be less well equipped in terms of infrastructure, use of resources and organizing capacity. To enforce a solution of development and achieve a good position, these cities have to aim on identifying their strengths and advantages obtained by using smart solutions.

By smarter, we understand the use smart solutions to make the world work efficiently. The smart city means to infuse intelligence in all activities for improve the quality of work and of life, to reduce cost and to improve the efficiencies.

We can define the concept of smart city like a new solution for a better manages of infrastructure and resources.

The term smart city is used in literature regarding the education of its citizens, the implying ICT in the production processes, the urban traffic and the inhabitants' mobility, the use of smart solutions in health or referred to the relation between the city government or administration and its citizens.

2. Necessity of smarter cities

Our society is characterized by urbanization – a large number of people live in our days in urban area; technological progress – in every day we can see new solutions for communication, transmission and storage of data; environmental changes – every activity in

our day is characterized by the important impact on natural resources how in fact are limited; economical growth – the gross domestic product of our world is substitute by the big cities how bring the people together and stimulate creativity and efficiency.

The entire elements who characterized our society drive us to implement a good manage of infrastructure, of resources and to carer to existing and future needs of citizens.

All the cities who implemented smart solutions had in view to improvement of citizen everyday life. In the last years by implementing smart solutions in different countries from European Union was made:

- Increasing the employment rate of employment for men and women aged between 20 and 64 years, while employing a larger number of young people, older and low-skilled people, coupled with a better integration of legal immigrants;
- Improving the conditions for research and development in order to increase investment levels and stimulate research, development and innovation of new indicators;
- Reduction of greenhouse gas emissions compared, increasing the share of renewable in final energy consumption and achieve increased energy efficiency;
- Improving education levels by reducing dropout rates and increasing the proportion of persons aged 30-34 years with university degrees or equivalent qualifications;
- Promoting social inclusion by reducing poverty and eliminating the risk of poverty.

The most significant advantages (Figure 1) are improved of citizen transportation, the access to city resources (libraries and public buildings, malls, networks etc.) and the opportunities for the employment and local growth (Dirks, 2010 and Barrientos, 2010).

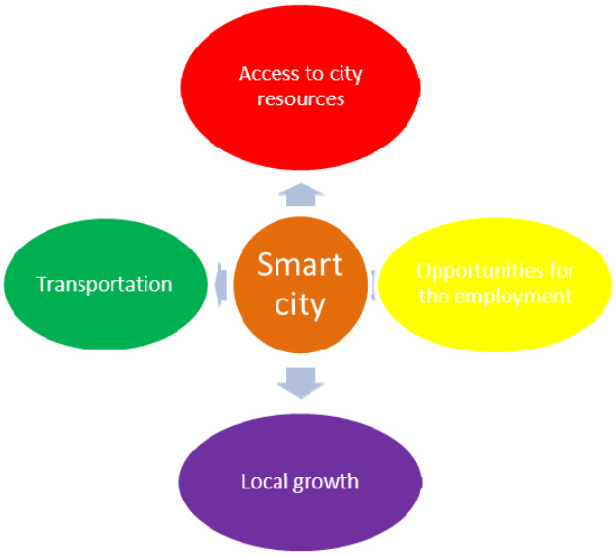

Figure 1.1 Smart city advantages

All these advantages highlight the need of implementations smart solutions in our country. The most important advantages and benefits of smart solutions are:

- use the last technologies;
- openness to smart technology;

- reducing costs;
- increasing degree of supports in case of disasters;
- improved the communication with citizens;
- citizen participation in community life economic, social and cultural;
- providing data coherent, consistent, updated;
- accessing open data;
- reducing time for solving citizens queries;
- reduce errors related to the information transmission security and transactions.

It is essential to highlight the necessity of smart solutions in city development. This can be described by a comparison between the traditional city and smart city (Table 1).

Table 1. Traditional city and smart city

Traditional city	Smart city
characterized by inefficiency of communication	better communication with citizens
inefficient use of resources	efficient use of resources
low access to administrative data	open access to administrative data
a lot of errors in information	reduce errors of information transmissions
without support in case of disasters	increasing the supports in case of disasters

These are only a part of elements who determinate us to try to implement a smart city. All these elements are the real support for urban development and for improve the quality of life.

It can take a lot of time for a city to become really smart. Sometime, the transformation is difficult to do because the mentality of citizens, or other time the evolutions can be stop by the natural disasters. But, to reduce costs, improve efficiencies, and deliver the quality of life citizens expect the implementing of smart city.

The urbanization without precedent of our cities and the technological progress on the hand, and on the other hand for a sustainable progress and for economic growth we need smart solutions for water, energy, transportation, healthcare, education, and safety or we can say we need smart solution for new cities.

3. Smart city characteristics

A question of our day is how do the cities smarter and what are the step and principles outlined above in the most cost-effective and productive fashion? The answer is to identify the high-impact areas of improvement of the city. These areas are correlated with the characteristics of the urban area. To the level of an urban city we can find seven characteristics (Figure 2). To describe a smart city and its seven characteristics it is necessary to develop a transparent and easy hierarchic structure, where each level is described by the results of the level below.

A set of factors were chosen to describe all the seven characteristics. For each characteristics we can identify a set of factors for describe the readiness and the use of smart solutions.

For smart economy were identified as factors: productivity and flexibility of the labour market, integration in the (inter-) national market, economic competitiveness, use of on-line solutions for sell or buy of products, quality of production, quantity of production, production diversity and research and development expenses of the authorities.

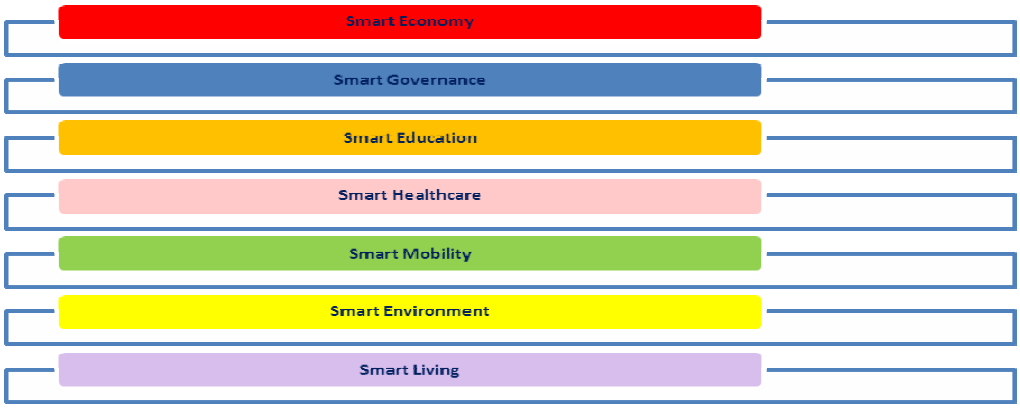

Figure 2. Smart city characteristics

Smart Governance includes factors like political participation, services for citizens as well as the functioning of the administration - we can say public and social services, transparent governance, political strategies and perspectives.

Smart Education is described by investment in education systems, education of the citizens, affinity to lifelong learning, education facilities, social interactions regarding integration and public life, research and investment in innovation and creativity.

Smart Healthcare is described by investment in healthcare systems, open access to medical health data, open access to patient data for interoperability of patient records, improved productivity of healthcare systems.

For smart mobility were identified as factors: local accessibility, (inter-)national accessibility, availability of ICT-infrastructure, sustainable, innovative and safe transport systems. In our days the smart cities promote walking, cycling, bike sharing, car sharing and smart mobility cards as part of an integrated mobility strategy to reduce travel times, fatal accidents and carbon emissions.

Smart environment is described by respected of natural conditions, environmental protection, level of pollution, sustainable resource management. And smart living cultural facilities, health conditions, individual safety, housing quality, education facilities, touristic facilities, social cohesion.

To analyze the performance of smart city a set of indicators can be assigned to each factor which describes one of the seven characteristics.

4. Smart city indicators

Observation of the economic and financial indicators illustrates the increases/ decreases in quantitative and qualitative. It is important to watch how ICTs are used for the sale and purchase products online products and changing turnover due to online sales.

- the number of ICT specialists;
- the number of electronic devices;
- the number of personal computers connected to the Internet;
- the number of Internet users;
- share of number of enterprises with Internet connection in all enterprises active
- investment in hardware;

- investment and expenses for information technology products and services;
- investments and expenditures of communication products and services;
- share of enterprises that have their own website in total company assets;
- share of enterprises that buy-sell business online in total assets;
- share of employees teleworking employees in total employees.

It is very important that each institution using the electronic system. There are some indicators that can inform us of the progress of computerization of government sector:

- ICT specialists;
- the number of electronic devices;
- the number of personal computers connected to the Internet;
- the number of Internet users;
- percentage of institutions with Internet connectivity in all institutions;
- share of investments and expenditures for ICT products and services in total investment and expenditure;
- investments and expenditures for ICT products.

Education sector includes both educational institutions in the public and private educational institutions. Also, the calculation of indicators will take account of all three levels of training and primary, secondary and tertiary (National Institute of Statistics classification).

- number of PCs per 100 students;
- number of PCs connected to the Internet per 100 pupils and students;
- Education institutions connected to the Internet in total number of educational institutions;
- The number of students using Internet.

By implementing the health sector of new electronic communications services will achieve an increase in efficiency, with lower costs. Using ICT in medicine results in economic growth as the healthcare system, as well as the patient fells.

- The number of equipment;
- The number of personal computers connected to the Internet;
- medical institutions share Internet connection in all institutions;
- Share of investments and medical expenses for ICT products and services in total investment and expenditure;
- Investments and expenditures for ICT products.

Analysis is needed ICT indicators for the household sector and influences the degree of development of settlements.

- Share of households with fixed telephone
- Share of households owning personal computers connected to the Internet in total households
- Share of total household expenditure communications

Observing indicators SME sector illustrates the increases / decreases in quantitative but also qualitative due to the implementation of new information and communication technologies. It is important to watch how ICT products are used to establish the long path that we must follow company.

- Share of enterprises that have personal computers;
- Share of employees using personal computers;
- Share of enterprises with Internet access;
- Share of number of employees using the Internet;
- Share of enterprises that Web site;
- Share of enterprises that sell via the Internet;
- Share of enterprises purchasing on the Internet.

Take account of all cultural institutions are a pillar for each locality harmonious development. Without this sector without investment in ICT products could be considered as the history and future of the relationship are not forgotten what contributes to the positive development of the village.

- share cultural institutions holding personal computers;
- share of employees using personal computers;
- share cultural institutions with Internet access;
- share of employees using the Internet;
- percentage of institutions of culture that Web site;
- share cultural institutions that sell over the Internet.

Make large investments in tourism revenues, which make this sector a pillar of economic development and the implementation of new information and communication technologies.

- the percentage of tourism that have personal computers;
- share of employees using personal computers;
- the percentage of tourism with Internet access;
- share of employees using the Internet;
- the percentage of tourism that Web site;
- the percentage of tourism that provides services via the Internet.

Analyses of the areas of a community are affected by the use of new information and communication technologies are support for achieving a smart urban area. It is essential to follow the use of advanced communication technologies in the local and among all sectors to assess the level of development of settlements upcoming smart counterpart.

To compare the different indicators it is necessary to use the same values. One method to standardize is by z-transformation:

$$z_i = \frac{x_i - \bar{x}}{s}$$

This method transforms all indicators into one value with an average 0 and a standard deviation 1. This method has the advantages to consider the heterogeneity within groups and maintain its metric information.

In (Dirks, 2010 and Barrientos, 2010) literature is highlight that we can implement new/intelligent solutions in our urban areas to become smarter and for this we can start from: transportation, education, healthcare and governance, which are the most important sectors of the city. In literature the attention is focus firstly on four high-impact areas of development.

5. Conclusions

The use of modern technology efficiently in our urban areas is an important part of modernization, growth and sustainable development. Our society today is more organized, smart and the information is situated in the center of it. In our age the cities development depend on the use of more and more smart solution. For sustainable growth of society is essential to use efficiently the modern technology and natural resources.

6. Acknowledgments

This work was supported from the European Social Fund through Sectored Operational Programmer Human Resources Development 2007- 2013, project number POSDRU/89/1.5/S/59184"Performance and excellence in postdoctoral research in Romanian economics science domain".

References

1. Barrientos, A. **A Planet of Smarter Cities: Security and critical infrastructures impact,** Retrieved from http://www.dintel.org/web/Eventos/CongresosEspana/Profesionales /2010/ponencias/barrientos.pdf
2. Chen-Ritzo, C.-H., Harrison, C., Paraszczak, J., Parr, F. **"Instrumenting the Planet"** IBM Journal of Research & Development. Vol. 53. No. 3. Paper 1. PP 1:16, 0018-8646, 2009
3. Dirks, S. , Gurdgiev, C. and Keeling, M. **IBM Global Business Services Executive Report**, **IBM Institute for Business Value Smarter cities for smarter growth "How cities can optimize their systems for the talent-based economy"**, IBM Corporation 2010, Retrieved from http://public.dhe.ibm.com/common/ssi/ecm/en/gbe03348usen/GBE03348USEN.PDF
4. Dirks, S., Keeling, M., Dencik, J. **IBM Global Business Services Government Executive Report, IBM Institute for Business Value How Smart is your city? Helping cities measure progress**, Copyright IBM Corporation 2009 Retrieved from http://public.dhe.ibm.com/common/ssi/ecm/en/gbe03248usen/GBE03248USEN.PDF
5. Dirks, S., Keeling, M. **IBM Global Business Services Government Executive Report IBM Institute for Business Value A vision of smarter cities How cities can lead the way into a prosperous and sustainable future**, Copyright IBM Corporation 2009, Retrieved from http://public.dhe.ibm.com/common/ssi/ecm/en/gbe03227usen/GBE03227USEN.PDF
6. Giffinger, R., Fertner, C., Kramar, H., Kalasek, R., Pichler-Milanović, N., Meijers, E. **Smart cities Ranking of European medium-sized cities, Final report**, October 2007, Research

Institute for Housing, Urban and Mobility Studies(OTB), Retrieved from http://www.smart-cities.eu/download/smart_cities_final_report.pdf

7. Palmisano, S. **Smarter Planet: The Next Leadership Agenda**, IBM Chairman and CEO to Council of Foreign Relations, New York City, November 6, 2008. Retrieved from http://www.ibm.com/ibm/ideasfromibm/ca/en/smartplanet/20090210/sjp_speech.shtml
8. **"IBM and Dubuque, Iowa Partener on Smarter City Initiative"**, 2009, Retrieved from http://www-03.ibm.com/press/us/en/pressrelease/28420.wss

---

[1] Lorena BATAGAN has graduated the Faculty of Cybernetics, Statistics and Economic Informatics in 2002. She has become teaching assistant in 2002. She has been university lecturer since 2009. She is university lecturer at faculty of Cybernetics, Statistics and Economic Informatics from Academy of Economic Studies. She holds a PhD degree in Economic Cybernetics and Statistics in 2007.
She is the author and co-author of 4 books and over 50 articles in journals and proceedings of national and international conferences, symposiums.