Comparing the diversity of information by word-of-mouth vs. web spread

ALON SELA[1(a)], LOUIS SHEKHTMAN[2], SHLOMO HAVLIN[2] and IRAD BEN-GAL[1]

[1] *Tel Aviv University - Haim Levanon St. 55, 6997801 Tel Aviv Israel*
[2] *Department of Physics, Bar Ilan University - Ramat Gan 52900, Israel*

Abstract –Many studies have explored spreading and diffusion through complex networks. The following study examines a specific case of spreading of opinions in modern society through two spreading schemes —defined as being either through "word of mouth" (WOM), or through online search engines (WEB). We apply both modelling and real experimental results and compare the opinions people adopt through an exposure to their friend's opinions, as opposed to the opinions they adopt when using a search engine based on the PageRank algorithm. A simulated study shows that when members in a population adopt decisions through the use of the WEB scheme, the population ends up with a few dominant views, while other views are barely expressed. In contrast, when members adopt decisions based on the WOM scheme, there is a far more diverse distribution of opinions in that population. The simulative results are further supported by an online experiment which finds that people searching information through a search engine end up with far more homogenous opinions as compared to those asking their friends.

Introduction. – Diffusion processes through complex networks have been studied in the context of disease epidemics [1–4] of computer viruses spread [4,5], as well as in the context of information spreading among people [6–16].

While many of the spreading models are general enough to provide insights on different spreading phenomena, such as detection of influential spreaders, system failures and influence of topologies [5,17], there are factors that are mainly relevant to information spreading among people.

In the context of individuals who adopt opinions, the choice is often among many opinions [18], unlike models for disease spreading [6,19] or spreading of computer viruses, where a node is either infected or uninfected.

Another unique factor to information spreading is that modern information spreading can occur via either physical or virtual interactions. In the process of a virus spreading, an infection tends to occur through the local interactions during a human-human encounter. This resembles information spread by word of mouth (WOM), where information diffuses only along the links of the network. Another common method, by which information

spreads globally throughout society, is through the internet (WEB) [20]. Such internet interactions are global and are often mitigated through a search engine.

An example of the type of decisions madethrough social influence is the choice of where to travel for vacation. In a network where influence occurs only through word of mouth, individuals search for information through their friends about their recent vacations recommendations, and then decide based upon the different suggestions received from their friends. In contrast, if an individual chooses to use the internet to look for a vacation location, he might use a search engine, which will provide him with the requested information.

Several previous works have studied the interactions between word of mouth and mass media [21,22], through Big Data meme tracking methods. Other works came to varying conclusions about the degree to which search engines based on PageRank-related algorithms [23], bias their search traffic results [11,24] and amplify the dominance of popular sites. However, none of these works considered the specific comparison, between the spread of ideas through search engines as compared to word of mouth.

(a)E-mail: `alonsela@tauex.tau.ac.il`

The present work develops an approach for studying modern information diffusion. It considers not only the biases in information flow resulting from the search engines' ranking algorithms, but also the bias which results from human behaviour and tendencies in the context of web searches. Such a bias can only be evaluated through a direct comparison between these two spreading mechanisms.

The spread of opinions by WOM and WEB have much in common. In both, a person is influenced by the opinions of his social network, and selects his personal opinion among these alternatives. In the virtual as in the digital worlds, social influence [10,25–28], will have a significant impact on the person's final decision.

The fundamental difference between the WOM spreading and the WEB spreading is that in the WOM the source of opinions is generated from real acquaintances while in the WEB it is from opinions fetched by an online search engine.

The simulated models for comparing spreading through WEB and WOM, find that information spread through WOM results in far more diversified opinions of the population. These results are further strengthened through an experimental study on real human subjects that supports these claims.

The proposed models. – The process of information spread can be divided into two stages: i) an *awareness stage* when a user only becomes aware of a new topic; and ii) an *evaluation stage*, when a user is exposed to opinions on the topic and has to select which opinion among these alternatives he/she would adopt.

While the awareness stage is similar in both the WEB and the WOM models, the evaluation stage is different. In the awareness stage, for both models, the user first becomes aware of the existence of a new topic by a neighbouring node which holds an opinion on the topic. After the user's first awareness, he collects information on the new topic through either WOM or through WEB methods. He then evaluates the information in order to form his own personal opinion. In the WOM model users search for information from their social connections, *e.g.*, family and friends, while in the WEB model they are exposed to opinions that are presented by the search engines following some online query.

For example, consider a user hearing his work colleagues talking about their locations for their summer vacation (awareness). The user might then seek information about a location for his own summer vacation. The user might search for such a location by asking his friends for their recommendations (WOM), or he can search for such using a search engine (WEB). The user will then evaluate among the options considered and decide for a vacation destination.

Information evaluation via WOM has been the subject of several studies [8,13,28–30]. In these studies, social influence is often modelled by the probability for adopting an opinion, which increases with the number of people holding this opinion in one's social circle. Similarly, the adoption of an opinion in the WEB is the outcome of similar social and cognitive processes. Thus, in general, the probability for adopting an opinion is proportional to its popularity, whether it is promoted by actual social connections or by web pages.

In the WEB, as in WOM, the probability for an adoption of an opinion increases as more web pages support this opinion. Apart from the fact that in the case of the WEB these opinions are collected by a search engine, and are written online, similar cognitive evaluation process are performed both for the WOM as for the WEB.

While the detailed algorithm for ranking pages by search engines is not fully known, PageRank is considered to be one of their most important aspects. The PageRank algorithm ranks well-connected web pages with higher grades, and presents highly graded web pages at the top of the search results list. In our WEB model we define the network as the network of users, *i.e.* the readers and the publishers of opinions on the internet. We assume that highly connected individuals publish their opinions in highly connected web pages. Accordingly, we calculate the PageRank score of the web page that publishes an opinion by the PageRank score of the person that holds this opinion. This score is then used to set the position of the opinion in the search engine result list.

The different opinions are first ordered according to the PageRank of their publisher, and then the searchers read these opinions. It is well known that the higher a web page appears in the search engine result list, the more likely it will be read by a searcher. This tendency is expressed through the Search Engine Result Page (SERP) function, which defines the probability of a person clicking on a link as a function of the relative position of that link in the search result list. The SERP function is a known probabilistic function that has been estimated from several surveys that are mainly performed by search engine optimization (SEO) firms [31,32]. We estimated the SERP function on the basis of 6 different surveys, which were conducted in the years 2006–2014.

The spread dynamics. Let $G = (V, E)$ be a social network of $|V| = N$ participants (nodes). At time $t = 0$, a small subset $V' \subseteq V$ of nodes is randomly chosen, and each node is seeded with one random opinion from all possible opinions $B = \{b_1, b_2, \ldots, b_l\}$. The spreading process then begins, using either i) WOM or ii) WEB spreading, and the opinion held by node i at time t is denoted o_i^t. Each user (node), is only able to read a limited number of k opinions from among the existing opinions. This limitation is especially important considering the vast amount of information available in the WEB which can never be fully read. In the WOM model, each opinions from the node's social circle might be adopted with a probability proportional to its rate of appearance in the social circle. In the WEB model, the probability of a node considering an opinion is derived by the SERP function. More precisely, in the WEB model, a list of all the opinions in the network is

first constructed and sorted by the PageRank of the node holding the opinion. Then, k opinions are chosen from this list as derived from the SERP function, and one of them is chosen. This process continues until all the nodes in the network adopt an opinion. Then, the final adoption fraction of each different opinion in the network is recorded while being sorted in descending order, to form a final adoption fractions R^{end}.

We note that the relative adoption fractions at late intermediate stages are found to be similar to the final adoption fraction where all the nodes accept an opinion.

After a node has adopted an opinion, a later change of opinion is not permitted in the current model. The rational for not allowing a change of opinion is the cost of opinion change. For example, cancelling a vacation after ticketing, can results in cancellation fees that would prevent (in most cases) such change of vacation destination after the act of conclusion has been made.

In the following, the spreading process is explicitly defined through WEB and WOM schemes.

The WOM spreading process

While not all nodes infected.
For each non-infected node u with at least one infected neighbour

1) *Create a list of the influencers opinions held by the neighbours of u that have an opinion (defined as IO).*

2) *Choose a random opinion from the list IO. Note that opinions present more often among the neighbours are more likely to be chosen.*

3) *Adopt the chosen opinion from step 2).*

Advance time in one time step.

The WEB spreading process

While not all nodes infected.
For each non-infected node u which has at least one infected neighbour

1) *Create a list of all the opinions of all the nodes in the network which have any opinion.*

2) *Sort the list by the PageRank of the node that holds the opinion (defined this list as AO).*

3) *Create from AO, a second list of k entries (opinions) which represent the actual opinions that would be read by an average user (denoted IO for influencers opinions). In the creation of IO from AO, the probability of reading an opinion located at position i in AO is given by the SERP probability function.*

4) *Choose a random opinion from the list IO.*

5) *Adopt the chosen opinion from step 4).*

Advance time in one time step.

Table 1: Summary of simulation parameters.

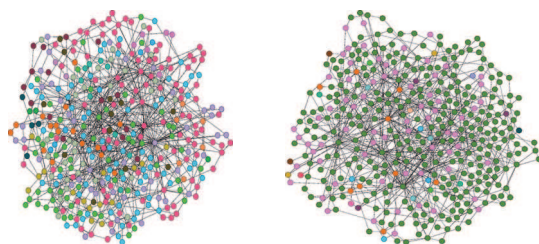| Denote	Parameter	Possible Values
PA	Preferential attachment	0, 0.5, 1
N	Num. of nodes	5000, 10000
m	Network density	2, 4, 8
k	Num. of opinions read	5, 10, 15
II	Num. of initial infections	15, 30, 45
S	Spreading model	WOM, WEB

Fig. 1: (Colour online) Illustrative comparison of simulations of spread of 15 opinions' in WOM (left) *vs.* WEB (right) in a small Barábasi-Albert network ($PA = 1$, $N = 200$, $m = 4$, $k = 10$, see parameters in table 1), where colours represent opinions adopted by the nodes. It can be seen qualitatively from the figure that the WEB spreading results in significantly less diversified opinions in comparison to the WOM spreading.

In the next section, we will present the simulation results, followed by results from an experiment with human subjects, which support the simulative results.

Results. –

Simulation parameters. The simulation set includes 8100 runs of opinions' spreading under different conditions and parameters, as indicated in table 1. Overall, in each simulation run, a network of size N was constructed, by implementing a preferential attachment process [5], in which each new node connects to m new nodes. The degree of preferential attachment process, denoted PA, varies with $PA = 1$ being a fully preferential attachment process, $PA = 0$ representing an Erdős-Rényi network, and $PA = 0.5$ being a process where in 50% of cases a random node is chosen, and in 50% of cases the selection is governed by a preferential attachment process.

For each combination of the parameters in table 1, 25 realizations were simulated, summing up in 8100 realizations overall. The vector R^{end} of the sorted final fractions of opinions' spread, for each of the 45 initially seeded ideas was then recorded. Several runs with networks of sizes $N = 20000$ and 30000 were inspected to verify the consistency with larger networks, but are not incorporated in the entire analysis due to their long running times by Agent Based Modelling.

Simulation results. We first present in fig. 1, a single representative realization for the WEB and WOM models
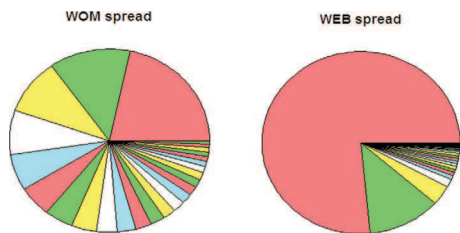
Fig. 2: (Colour online) Final average fractions of adoption for spread of different opinions, as generated by the WOM and the WEB simulations when using table 1 parameters. Note that the WOM model results in a significant higher variability of opinions' spread in the population.
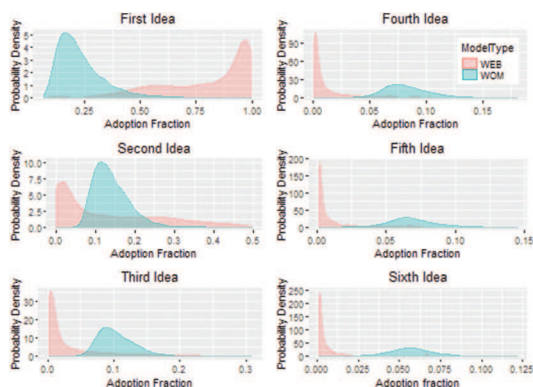

Fig. 3: (Colour online) Adoption fractions of six top ideas in the network, sorted by their popularity (the 1st idea is the most common one) for both the WOM (marked by azure) and the WEB (marked by red) models. While the most common idea spreads to over 80% of the network by the WEB, the less common ideas are observed in WOM but are barely noticeable in WEB.

which is added for descriptive purposes of the main results of the paper. The averages of the spreads fractions are then presented in fig. 2 and in fig. 3, which demonstrate the higher variance in the WOM model as compared to the WEB model. The difference between the two models is first presented in fig. 2, where the most common idea is adopted by approximately 75% of the nodes in the WEB spread, but only by 23% of the nodes in the WOM spread. In the WEB spread however, over 95% of the nodes adopted on average only three ideas, while in the WOM, 95% of the nodes adopted as much as 15 ideas.

A comparison of the fractions of populations which adopted the 1st most popular idea, the 2nd, ..., 6th most popular ideas is presented in fig. 3. It reveals that the adoption fraction of the 1st most popular idea for the WEB spreading model (red histogram) is significantly larger on average than that for the WOM spreading model. The peak in the WEB (red) histogram for the 1st idea occurs between the values of 0.85 and 1, and is mainly the outcome of simulations runs where one single idea is adopted by a large fraction of nodes. The dominance of

one single idea in the WOM spread is far less drastic. In comparison, the 1st idea in the WOM spreading model (azure histogram), has a lower mean adoption rate of approximately 0.23, and follows a narrower Gaussian distribution. This trend flips, from the 2nd most popular idea onward, where the mean of the WOM model is larger than the mean of the WEB model. Thus, when comparing the adoption fractions in the 5th and, 6th popular ideas, in the WOM model these ideas still capture a reasonable fraction of the population, whereas in the WEB model these ideas have barely spread at all.

It should be noticed that while on average this flipping occurs at the second idea, it may occur at a different point. It might be possible that if the SERP click through rate function decayed less steeply, *i.e.* people would click more on later results; the flip would occur at later ideas. However, given the current estimate for the SERP click through rate function, the dominance of the first idea in the WEB spread is such that on average the second idea is already more popular in the WOM than in the WEB.

The differences in the diversities between the two models are also evaluated by comparing the average entropy of the opinion spread in each model. This difference is found significant (t-test, p-value 2.2E-16), with the mean for the WOM model being 3.608, while for the WEB it was only 0.8337, respectively.

Experimental results with human subjects. To test our conclusion that the use of the WEB method results in more homogenous opinions in the population, we conducted an experiment, based on real human subjects.

Two groups of users were required to answer the same set of questions. One group was requested to search the answer solely by using Google, while the other was instructed to answer the questions by asking their friends but without any search engine. The three questions were:

1) *What is the best new car to buy?*

2) *What is the best country for a vacation overseas?*

3) *What is the best restaurant in New York?*

These three questions were answered by 100 responders, which were gathered from Amazon Mechanical Turk service, a crowdsourcing marketplace operated by Amazon, which enables recruiting workers for simple and repetitive tasks.

From these 100 responders, each answering all three questions, 50 used the WEB scheme while the other 50 used the WOM scheme.

After cleaning the data and combining similar answers such as "London" and "England" in Question 1, the final results included 49 WEB responders and 49 WOM responders, each of whom answered all three questions and a total of 294 complete answers have been reported.

For example, UK was repeatedly indicated as the best location for vacation in 26/49 responses (53%) among the

Table 2: Answers to 3 questions.

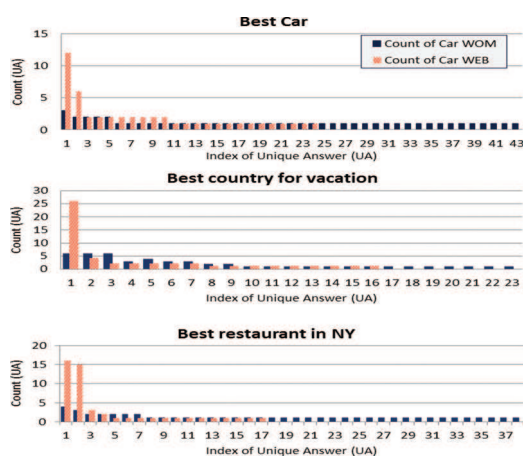| Question 1: Best car	WEB	WOM
Num. of different uniqe answers	24	43
Repeats of most common answer	12	3
Question 2: Country for vacation		
Num. of different uniqe answers	16	23
Repeats of most common answer	26	6
Question 3: Best restaurant in NY		
Num. of different uniqe answers	17	38
Repeats of most common answer	16	4

Fig. 4: (Colour online) The distribution of answers for the three questions ("best car", "best country for vacation", "best restaurant in NY") as obtained by the WOM search (azure) *vs.* the WEB search (red). The *x*-axis is the answer index, and the *y*-axis is its number of reoccurrences.
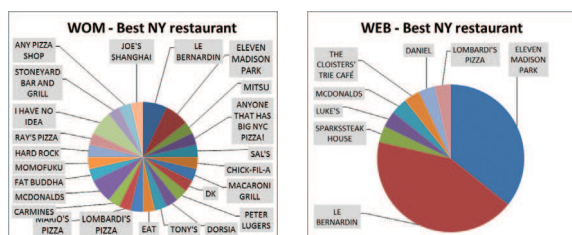

Fig. 5: (Colour online) Information diversity for the question "best restaurant in NY" as received from users using WOM *vs.* WEB searching methods.

WEB users, while Australia and Japan were most popular in the WOM model with only 6/49 users (12%).

Furthermore, as can be seen in table 2, while the WEB model resulted in 17 different opinions for the "best restaurant" question, and as much as 16 responders repeating the same restaurant as the best restaurant in NY, the WOM model included as many as 38 different

"best restaurant" answers with only 4 repeating names of restaurants.

The experimental results strongly support the model simulation since all three questions included a lower variability of information while using the WEB as compared to the WOM. These results can be seen in fig. 4, while the most extreme reduction in the diversity of information is seen in the NY restaurants question which is presented in depth in fig. 5.

Conclusion. – Our results suggest that the use of WEB search engines substantially decreases the diversity of opinions in a population, compared to word-of-mouth (WOM) spreading. While previous studies have attempted to suggest that web search results are less biased than believed [23] and that the distribution of internet pages is less unbalanced than expected, we suggest that users' decisions are still highly biased when using the WEB search engine since each user ends up reading similar opinions for similar searches. This is the result of two independent "rich get richer" processes, where the first is found in the search engine algorithm and the second is found in users' behaviour as expressed in the SERP function. Such similarity in the exposure to opinions might sometimes lead users to make similar decisions and thus increases homogeneity in the population.

In certain cases the homogeneity of the WEB may be desirable, particularly in cases where there is a clear optimal choice. The wisdom of the crowd is often a powerful tool in helping groups arrive at best decisions. In other cases, however, diversity of opinions may be preferable and personal tastes can play an important role. Particularly in the case of personal tastes, it may be more useful to ask friends' opinions (WOM) since they have more knowledge of what an individual will like or dislike, while answers on the WEB have no such knowledge. Also, diversity of opinions is known to have its advantages in creative processes [33,34]. In cases where a diversity of opinions is required, this work recommends to include (at least to some degree), the WOM information search and spread, which can be obtained by attending conferences or using social networks which are seen as WOM information search. This is particularly important as people rely more and more on search engines. Measurement of the influence of solely using search engines to search information, in creative processes, can be a subject for further future research.

REFERENCES

[1] Newman M. E., *Phys. Rev. E*, **66** (2002) 016128.
[2] Brockmann D. and Helbing D., *Science*, **342** (2013) 1337.
[3] Pastor-Satorras R. and Vespignani A., *Phys. Rev. Lett.*, **86** (2001) 3200.
[4] Gallos L. K. and Argyrakis P., *Physica A: Stat. Mech. Appl.*, **330** (2003) 117.

[5] BARABASI A.-L. and ALBERT R., *Science*, **286** (1999) 509.

[6] PARSHANI R., CARMI S. and HAVLIN S., *Phys. Rev. Lett.*, **104** (2010) 258701.

[7] BANERJEE A., CHANDRASEKHAR A. G., DUFLO E. and JACKSON M. O., *Science*, **341** (2013) 1236498.

[8] DOMINGOS P. and RICHARDSON M., *Mining the network value of customers*, in *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2001.

[9] EASLEY D. and KLEINBERG J., *Networks, Crowds, and Markets: Reasoning about a Highly Connected World* (Cambridge University Press) 2010.

[10] GRANOVETTER M., *Am. J. Sociol.*, (1978) 1420.

[11] PENNOCK D. M., FLAKE G. W., LAWRENCE S., GLOVER E. J. and GILES C. L., *Proc. Natl. Acad. Sci. U.S.A.*, **99** (2002) 5207.

[12] WATTS D. J., *Proc. Natl. Acad. Sci. U.S.A.*, **99** (2002) 5766.

[13] KEMPE D., KLEINBERG J. and TARDOS E., *Maximizing the spread of influence through a social network*, in *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2003.

[14] FENG L., HU Y., LI B., STANLEY H. E., HAVLIN S. and BRAUNSTEIN L. A., *PloS ONE*, **10** (2015) e0126090.

[15] RADICCHI F., FORTUNATO S. and VESPIGNANI A., *Citation Networks*, in *Models of Science Dynamics* (Springer) 2012, pp. 233–257

[16] STEHL'E J., BARRAT A. and BIANCONI G., *Phys. Rev. E*, **81** (2010) 035101.

[17] CHEN D., LU L., SHANG M.-S., ZHANG Y.-C. and ZHOU T., *Physica A: Stat. Mech. Appl.*, **391** (2012) 1777.

[18] WENG L., FLAMMINI A., VESPIGNANI A. and MENCZER F., *Sci. Rep.*, **2** (2012) 335.

[19] KERMACK W. and ANDERSON M., *Proc. R. Soc. London*, **115** (1927) 700.

[20] RYBSKI D., BULDYREV S. V., HAVLIN S., LILJEROS F. and MAKSE H. A., *Proc. Natl. Acad. Sci.*, **106** (2009) 12640.

[21] LESKOVEC J., BACKSTROM L. and KLEINBERG J., *Meme-tracking and the dynamics of the news cycle*, in *of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM) 2009, pp. 497–506.

[22] GOMEZ RODRIGUEZ M., LESKOVEC J. and KRAUSE A., *Inferring Networks of Diffusion and Influence*, in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM) 2010, pp. 1019–1028.

[23] FORTUNATO S., FLAMMINI A., MENCZER F. and VESPIGNANI A., *Proc. Natl. Acad. Sci. U.S.A.*, **103** (2006) 12684.

[24] PAGE L., BRIN S., MOTWANI R. and WINOGRAD T., *The PageRank Citation Ranking: Bringing Order to the Web*, Technical Report, Stanford InfoLab (1999).

[25] BOND R. and SMITH P. B., *Psychol. Bull.*, **119** (1996) 111.

[26] WATTS D. J. and DODDS P. S., *J. Consum. Res.*, **34** (2007) 441.

[27] ZHAO K., STEHL'E J., BIANCONI G. and BARRAT A., *Phys. Rev. E*, **83** (2011) 056109.

[28] RICHARDSON M. and DOMINGOS P., *Mining Knowledge-Sharing Sites for Viral Marketing*, in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM) 2002, pp. 61–70.

[29] JANSEN B. and SPINK A., *The Effect on Click-Through of Combining Sponsored and Non-Sponsored Search Engine Results in a Single Listing*, in *Proceedings of the 2007 Workshop on Sponsored Search Auctions, WWW Conference*, 2007.

[30] DE BRUYN A. and LILIEN G. L., *Int. J. Res. Mark.*, **25** (2008) 151.

[31] PAN B., *Tour. Manag.*, **47** (2015) 79.

[32] PETRESCU P., *Google Organic Click-Through Rates in 2014, Moz, 1 October 2014* (Online), `https://moz.com/blog/google-organic-click-through-rates-in-2014`.

[33] KASPERSON C. J., *Human Commun. Res.*, **4** (1978) 113.

[34] PAULUS P., *Appl. Psychol.*, **49** (2000) 237.