Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Purchasing & Supply Management 10 (2004) 211–222
JOURNAL OF PURCHASING AND SUPPLY MANAGEMENT
www.elsevier.com/locate/pursup

The myth of purchasing professionals' expertise. More evidence on whether computers can make better procurement decisions

Frits Tazelaar[a], Chris Snijders[b,*]

[a]Department of Sociology/ICS, Utrecht University, Heidelberglaan 2, 3584 CS Utrecht, The Netherlands
[b]Department of Technology Management/ECIS, Subdept. Of Technology and Policy, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands

Received 28 June 2004; received in revised form 1 September 2004; accepted 20 November 2004

Abstract

In a previous experiment, we have shown that risk assessments of purchasing experts are certainly *not* better than that of subjects untrained in purchasing, and worse than the decisions made by formal models (J. Purchas. Supply Manage. 9 (2003) 191–198). Since both these results are rather counterintuitive, we conducted a series of experiments geared at replication and extension of these findings. These new experiments show that our previous results are robust, and reveal an additional finding that is both worrying and puzzling. It actually seems to be the case that for the purchasing decision tasks in our experiments, experts perform worse with growing experience. It therefore seems that, at least for the kinds of purchasing decisions under study, it does not make much sense to use expert judgments at all. However, we show that there is a way in which expert judgments can be used in combination with formal models to improve the predictive accuracy of purchasing predictions. In our case, superior predictions are made when we combine the prediction of a formal model with the prediction of the 'average expert', thereby combining the robust linear trends as encapsulated in the formal model with the more intuitive configural rules used by experts. We provide several explanations for this phenomenon.

1. Introduction

Most people would agree that at least one of the tasks of a purchase manager is to decide which of a set of purchasing transactions needs purchase management more. For instance, for some transactions it makes sense to ask for many tenders, invest much in the screening of suppliers, involve much time in negotiating, and put a serious effort in writing a detailed contract. For other transactions such investments are not effective or not efficient (Batenburg et al., 2000).

Although it is typically part of a purchase manager's job, there are compelling arguments on the basis of the literature on *clinical* versus *statistical prediction* that suggest that purchase managers—like all other humans—are typically not good at making precisely these kinds of judgments. In a review study, Grove et al. (2000) have shown that for single, quantitative decision tasks computer models almost always perform at least as good as or better than human experts. Most of the studies they reviewed were based on tests in medical, forensic and clinical-personality studies (102 out of 136). There are only a handful of studies comparing human experts with models that deals with 'more economic topics'. Grove et al. (2000, Table 1) mention studies on business failure, job performance, job turnover, business startup success, job success and work productivity. In a previous publication (Snijders et al., 2003), we set out to test this assertion when it comes to judgment and decision-making in purchasing, and reported on an alarming and somewhat counterintuitive result. It indeed turned out that, for the cases under study,

*Corresponding author. Tel.: +31 40 247 2640; fax: +31 40 246 4646.
E-mail address: c.c.p.snijders@tm.tue.nl (C. Snijders).

1478-4092/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.pursup.2004.11.004

Table 1
Spearman correlations between actual and predicted scores, averaged per group, as in Snijders et al. (2003)

Formula	0.37
Students	0.26
Purchasing professionals	0.24

purchasing professionals do *not* make better purchase decisions than undergraduates, and both are actually outperformed by a computer model (using a simple formula).

Though in principle this result is in line with research in other areas, and in that sense perhaps need not be treated as a surprising finding, we want to analyze some of the evident follow-up questions related to this result. First, we consider how robust this finding is. Perhaps our finding was simply for a purchasing professional rather unlucky statistical fluke, so we tried to replicate our findings (and succeeded). Second, we extended our previous research and considered possible reasons for our findings by looking at the way in which experts, in general, tend to behave. Finally, we focus on whether and how the decision-making of purchasing professionals can be improved, based on the literature on experts and expertise.

2. Computer beats purchasing professionals: a robust finding

We first repeat in brief the essential elements of the experimental setup as used in Snijders et al. (2003). In that experiment, both purchase professionals and undergraduates were each given 8 case descriptions regarding a procurement transaction (the procurement of IT-products; see Fig. 1 for an example case), and were asked, among other things, to predict the likelihood of this transaction being a problematic one (see Snijders et al., 2003 for details). All our claims about the judgment and decision-making capabilities of purchasing professionals are, therefore, based on purchasing professionals being able to identify which transactions are the ones most likely to be problematic. This is, obviously, not the only kind of decision a purchasing manager has to make, but we think it is an important one. Moreover, the purchasing professionals themselves felt they would do fine in this task, even when asked after completion (but before displaying the results).

In fact, the vignettes were chosen from a larger database of real purchasing transactions, so that we actually knew the correct answers to what the purchasing professional and the students were predicting and could compare their answers with the real ones (for a more detailed description of this database, see the Appendix A, and Batenburg, 1996 or Buskens and Batenburg, 2000). In short, our professionals and students were to predict how many problems would occur, and we knew the right answer. All undergraduates involved were freshmen in information sciences and participated as part of a course requirement. Purchasing professionals participated in reaction to an invitation from a student. Each pair of students had to find one purchasing professional who was willing to participate. Preferably, the professional should have experience in both purchasing and IT, since the transactions on the vignettes were all about IT-products. Ultimately, 30 purchasing professionals and 60 students participated.

We made sure that for (almost) all sets of eight vignettes, there were three individuals who made predictions: a purchasing professional and two students. The 240 (30×8) vignettes given to the purchasing professional were all different. This guaranteed a large spread in the kinds of vignettes under consideration and it enabled a clean comparison of purchasing professionals versus students. Participants were also asked to answer several other questions to which we will return later.

Beforehand, we calculated a formula that generates predictions on the likelihood of a particular transaction being problematic (prediction 1) and which kinds of problems were to be expected (prediction 2, for four kinds of problems). No fancy modeling was used: for both issues the formula was linear in the predictors. For instance, the formula to predict the likelihood of problems was a linear combination of the price of the product (if high, then more problems), the importance of the product for the profit of the buyer (if high, then more problems), the buyer's ability to judge the price/quality ratio of the product (if high, then more problems), and the degree of detail in the written contract (if high, then more problems). Note that this , roughly shows that the model simply expects more problems for larger and more complicated transactions, irrespective of the amount of effort involved in careful planning and the kind of partner.[1] We did not include the kind and number of products as a predictor for the model; this even gives a small information deficit to the model as compared to the respondent.

Hence, we used five formulas: one to predict how problematic a transaction was going to be, and four separate formulas, one for each specific problem that could occur. Since we calculated the formulas on the

[1] More careful and elaborate measurement and data analysis using the dataset as mentioned in Buskens and Batenburg (2000) show that other aspects than just these four also correlate with the amount and likelihood of problems in a purchasing transaction (cf. Rooks, 2002). That the model to predict problems as used in this paper is relatively straightforward is in part a consequence of the fact that one only has 14 separate items available for prediction.

The product

It concerns the purchase of	<pc(s),cabling, tailor-made software>
Price in Dutch guilders (Euros in later studies)	< between 100,000 and 200,000 >
Importance for the profit of the buyer firm	< high >

The supplier

Size of the supplier (number of employees)	< 400 >
Reputation in the market	< reasonable >

The buyer

Number of employees	< 35 >
Number of years in business with same supplier	< 2 >
Purchasing arranged through	< purchasing department >
Legal issues arranged through	< external experts >
Can judge price/quality for the different possible suppliers	< hardly >
Know other clients of the same supplier	< yes >

Ex ante purchasing management

Number of tender	< 2 >
Total investment in search, screening, selection, Negotiating, and contracting	< 4 mandays >
Degree of detail of the written contract	< high >

On a scale from 0 to 100:
How problematic do you think this transaction will turn out to be? _____

where: 0 = completely unproblematic
100 = highly problematic

Which of the following problems do you expect to have a high probability of occurrence for this transaction? (you may choose more than one)

0 late delivery
0 over price / budget
0 incorrect specifications upon delivery
0 inadequate documentation

How certain are you about your answers?

"I am just guessing"	1	2	3	4	5	6	7 "I am absolutely certain"

Fig. 1. Example vignette. Words between 〈 〉 varied across vignettes.

basis of data that we had set aside for this purpose and had *not* used in the experiment, this ensures standard cross-validation of our formula.

In this setup, a comparison of both the answers as calculated from the formula and the answers as provided by the participants with the actual answers enabled us to say something about whether computers can outperform humans in predicting (the probability of) problems associated with IT-transactions. The comparison runs as follows. For each participant we calculated the spearman rank correlation between the predicted scores and the actual scores, across the eight cases. A subject who ordered the vignettes perfectly would get the maximum score of +1, whereas ordering the transactions precisely the wrong way around yielded the minimum score of –1.[2] We then compared the average scores across the groups of participants: undergraduates, purchasing professionals, and the formula. Further details can be found in Appendix A and in Snijders et al. (2003).

The results were that the purchasing professionals performed worst, even a (non-significant) bit worse than undergraduates. Our formula outperformed both groups (Snijders et al., 2003: p. 195).

[2]There are several other sensible ways to calculate measures indicating how adequate the predictions of humans and the model are. For instance, one could use a Pearson correlation instead of the reported rank correlation, or focus on the absolute differences per case instead of calculating a measured based on the set of eight vignettes. Our findings are robust to such different ways of measurement.

Table 2
Spearman correlations between actual and predicted scores, averaged per group, across experiments (higher is better)

	Original	I&L-1	I&L-2	MC	Rep03
Formula	0.37	0.29	0.31	0.28	0.34
Non-professionals	0.26	—	—	—	0.14
Professionals[a]	0.24	0.15	0.22	0.17	0.17
# participants	91	72	72	13	118

[a]For I&L-1 and I&L-2 we grouped together all participants, even though not all of them label themselves as strictly a *purchasing manager*, since all participants (with the exception of just one or two) have some professional relation to purchasing.

To assess the robustness of this finding, we ran (almost) the same experiment in several ways.

The I&L test: In March 2003, Paul van Haaster, the editor of the Dutch "Tijdschrift voor Inkoop en Logistiek" (I&L; "Journal of Purchasing and Logistics"), a management journal of the NEderlandse Vereniging voor Inkoopmanagement (NEVI; Dutch Association for Purchasing Management) called forth to the readers to participate in our "I&L Purchasing Test". Readers were informed that we believed that our formula could outperform the readers with respect to predicting the likelihood of problems, and were directed to our website if they wanted to participate to see if we were right (www.science2business.nl). On our website, we basically asked the participants, most but not all of them purchasing managers, to do the same test as described above. Again each participant received eight cases sampled from our database, for which they had to predict the likelihood of the case being a problematic one. After about 6 weeks, the official test was closed and the results were reported in I&L (Snijders, 2003). We label this experiment I&L-1. After the publication of the results in I&L, people could still participate on the website and, indeed, some more participants completed the test since then. To discern these data from the original test, we label these results I&L-2. Participants were allowed to do the test more than once, and some did. We report the results based on respondents' first runs of eight cases only.

The masterclass: In January 2002, we gave the same test to 13 participants of an International MasterClass in Strategic Purchasing and Supply Management, at Corsendonk, Belgium (MC). All participants, from France, Belgium, The Netherlands and the United Kingdom, were professional experts in the field of purchasing.

Replication and extension: In November 2002, we replicated our original experiment (including a few extensions) in the same university course in which the original experiment was conducted (Rep03). Next to purchasing professionals and students 'super laymen' also were invited to participate in the experiment: people who were specifically chosen *not* to have any specific knowledge about purchasing, or about IT-products.

The results of these experiments are reported in Table 2.

In all experiments, the formula performs best. We take this as rather strong support for our claim that, at least in such a setting, a computerized prediction outperforms humans, even if these humans are professionals in the field of purchasing.

As we outlined earlier, although purchasing professionals—or even purchasing academics—might frown at the credibility of these results, the results are actually not that surprising in light of previous research in other areas. Formulas are often found to predict at least as good or better than experts (Meehl, 1954, 1986; Dawes, 1971, 1979; Kleinmuntz, 1990; Dawes et al., 1993; Grove and Meehl, 1996; Grove et al., 2000; Snijders et al., 2003). One likely reason for this, as often mentioned in the literature, could be that humans in general are not that good at tasks where sound decision-making involves reliably storing, retrieving, and combining information. But why is it that this also holds for professionals (in purchasing)? One would expect that the professionals have learned to overcome the problems associated with making accurate predictions based on their expertise and experience. We set out to better understand why purchasing professionals are performing as they do. Part of the answer lies in the research on expertise in general, from which we first highlight some important findings.

3. Previous research on expertise: how experts are different

For a relatively long period of time, the literature on expertise has been surprisingly unspecific on what constitutes an expert in a field. In part, this might have come about because research in cognitive psychology, typically the academic discipline that is interested in the area of expertise, studied "obvious experts". For instance, expert decision-making has been studied quite extensively in chess, where grandmasters constitute the perfect example of an expert because a chess player's ELO-rating is an accurate measurement of a player's strength: find a grandmaster based on ELO-rating (say,

2600+) and you have found an expert. As the study of expert decision-making progressed, and more and more other academic disciplines became involved in this topic, one has become a bit careless in what it means "to be an expert". Usually, an expert is simply a person who is experienced in making predictions in a given domain and has some professional or social credentials (cf. Camerer and Johnson, 1991: p. 196). Usually, the minimum condition to label someone an expert is the possession of some specific human capital; both training and experience in a specific domain of expertise. The problem is that if we are looking for professional experts in purchasing, or extraordinary experts in management, we have no clear and objective measurements and must rely on cues other than measurable external validity. This is nicely illustrated in Shanteau (1987) and Shanteau and Peters (1989). Shanteau proposes a set of characteristics that determines whether and to what extent someone must be considered an expert, divided in knowledge and cognition, personality traits, and presentation and image (see Table 3).

As can be seen from Shanteau's list, to a substantial extent the cues that are used are not necessarily related to being objectively able to perform better than non-experts. For instance, self-confidence need not be related to true expertise at all, just like communication skills. Likewise, being able to quickly and seemingly effortlessly make decisions does not guarantee that these decisions are accurate. In fact, the list is closer to a list of necessary conditions. If you really are an expert, then you are likely to score high on this list. Whether the reverse also holds, is not obvious, and often not true (Camerer and Johnson, 1991). In the tradition of the field, the list largely represents process characteristics of experts: ways in which the experts deal with information.

A similar set of expert characteristics is given in Glaser and Chi (1988). In a review, they conclude that experts differ from non-experts in the following ways (Table 4):

For the larger part, Glaser and Chi's conclusions are based on research in "beta-oriented" areas, such as physics (and chess), so it remains to be seen to what extent these conclusions carry over to purchasing. Nevertheless, it does show that documented differences between experts and others exist. It is useful to emphasize this, to preclude that one would misinterpret the message of this paper as "experts know and can do nothing". This is certainly not what we are claiming. What is noteworthy though, is that again a large share of the characteristics relate to the *way in which* experts process information and (can) make decisions: they see

Table 3
Generic characteristics of experts (summarizing Shanteau and Peters, 1989)

Knowledge and cognition	
An extensive knowledge base	Experts have an extensive knowledge base and make a serious effort to keep up with the current facts, trends, and developments
Creativity	The ability to provide novel or even unique solutions to difficult problems. They are capable of generating new approaches to established problems as necessary.
Perceptive	The ability to extract information from a problem that others cannot see. Experts' decision-making ability is enhanced by insightful recognition and evaluation of confusing situations.
Knowing what is relevant	On the basis of experience, experts can readily distinguish relevant from irrelevant information in a problem. They utilize only what is relevant; the ignore what is not.
Simplification	Expert know how to use a divide-and-conquer approach with complex problems. They work on parts to get a better understanding.
Identifying exceptions	Experts know when to follow established decision strategies and when not to. They don't have just one way to solve problems.
Personality	
Self-confidence	Experts have a strong belief in their ability to make good decisions. They are calm and self-assured while making decisions.
Adaptability	Experts adjust their decision-making strategies to fit the current situation. They are responsive to changes in conditions of the on-going problem.
Experience	Experts use past experience to make decisions more-or-less automatically. Their background and experience produces decisions without obvious effort.
Stress Tolerance	Experts are able to make decisions under high stress situations. They continue to be effective problem solvers even as conditions progressively worsen because of high levels of pressure.
Presentation/image	
Communication Expertise	Experts can convince others that they have specialized knowledge. They can effectively communicate their ability to make decisions to others.
Problem Selectivity	Experts use foresight and planning in selecting which problems to work on and which not. They tackle those problems that they can effectively handle or solve.
Assumes Responsibility	Experts accept responsibility for the outcomes of decisions, successful or unsuccessful. They are willing to stand behind their decisions.

Table 4
Expert characteristics

1. Experts excel only in their own domain (expertise is narrow)
2. Experts consider larger coherent pieces (“chunks”) of information
3. Experts are quicker
4. Experts have a better memory (in their own domain)
5. Experts understand the underlying problem at a deeper level
6. Experts spend a relatively large part of their time analyzing a problem qualitatively (“what should roughly be the outcome?”, “to which category of problems does this problem belong?”)

Source: Glaser and Chi (1988).

and consider larger chunks of information, they are faster, decide on the basis of a better memory, and so on. This suggests that experts are better equipped to make decisions. Naturally, a better memory and higher speed are assets, but they do not necessarily imply that those who possess these qualities actually make better decisions. The *process-performance paradox* (Camerer and Johnson, 1991) is that in many fields ‘experts’ can often indeed be shown to behave and decide differently, but often not with better objective results. It certainly seems that our purchasing professionals fit this description: they do not perform better than non-experts. Let us find out if there are differences in the way in which the experts make their decisions in our experiment.

4. Purchasing professionals versus laymen

In all experiments we asked the participants after the completion of the cases if they could divide 100 points over the 14 different case characteristics. They should give more points to characteristics they considered “more important to predict the degree of problems”. For instance, if a participant thought that price is the only thing that matters to predict the degree of problems, this participant should give all 100 points to the aspect “price”. Or, if a participant thought that mainly the degree of detail in the written contract matters, together with the reputation of the supplier, then “degree of detail” could get, say, 70 points, and “reputation” 30 points. Table 5 reports the median scores per case characteristic, distinguishing between purchasing professionals and others (we chose medians instead of means because these are less sensitive to outliers).

There is certainly some agreement here: both purchasing professionals and others feel that the degree of detail of the written contract, and the kind and number of the products, are important. Many other case characteristics are deemed equally important by both groups of participants. There are also some differences. Professionals feel reputation is less important, whereas the non-professionals think it is one of the most important aspects. Instead, professionals attach predictive value to whether the buyer can judge price/quality, about which the non-professionals do not feel that strong.

Table 5
Median importance of the different case characteristics in predicting how problematic a case will be, according to the participants. Data pooled across experiments, distinguishing between purchasing professionals and others. Numbers are rounded to the nearest integer

	Purchasing pro	Others
The product		
Kind and number of products	10	9
Price	5	5
Importance for the profit of the buyer firm	5	5
The supplier		
Size of the supplier (number of employees)	5	4
Reputation in the market	6	10
The buyer		
Number of employees	1	3
Number of years in business with same supplier	5	9
Way in which purchasing is arranged	5	5
Way in which legal issues are arranged	4	5
Can judge price/quality for the different possible suppliers	10	7
Know other clients of the same supplier	2	5
Ex ante purchasing management		
Number of tenders	7	5
Total investment in search, screening, selection, negotiating, and contracting	6	7
Degree of detail of the written contract	10	10

Interestingly, there is some evidence that the professionals are indeed “chunkier” in their way of dealing with this division of points. For instance, the professionals show a larger standard deviation in their way of dividing points (note this cannot be inferred from Table 5). Their standard deviation across the fourteen characteristics is 7.4 whereas it is 6.4 for the non-professionals. Moreover, if we count the number of characteristics to which professionals give points at all, then the average number is 10.0 for the professionals and 11.2 for the non-professionals. We now look at the way of dividing the points in some more detail.

For one experiment, IL-1, we were able to distinguish three characteristics on which the professionals typically differ. The first one is whether the professional is a person for whom purchasing is their core business versus the ones for whom it is not (*core business*). The second characteristic distinguishes purchasing professionals with a lot of experience versus those with only a brief history in purchasing (*experience*). The third characteristic differentiates between purchasing professionals

with a relatively "high-level job" (senior purchasers, heads of departments, heads of purchasing) versus those with jobs that are more at the operational level (*level*).

Comparing the test performance of each of the distinguished categories results in some noteworthy findings. First, purchasing professionals for whom purchasing is their core business hardly perform better (Spearman correlations of 0.20 versus 0.13, n.s., $p > 0.20$). Second, purchasing professionals with a lot of experience in purchasing are not better than purchasing professionals with only a brief history in purchasing. On the contrary, experienced purchasing professionals perform (significantly) worse. This finding is even more remarkable, given the fact that the experienced professionals have more trust in their own abilities to perform well on such a test: the more years in purchasing, the more purchasing managers *claim* that they can judge the likelihood of problems in purchasing transactions on the basis of circa ten transaction characteristics (Fig. 2).

Third, purchasing professionals with a relatively "high-level job" do not perform any better in judging the likelihood of problems in purchasing transactions than purchasing professionals with jobs that are more at the operational level. On the contrary, it turns out that senior purchasers, heads of departments, and purchasing directors perform (significantly) worse. They score an average Spearman correlation of 0.01—which implies that their decisions cannot be distinguished from flipping a coin—versus 0.29 for the juniors.

These results put a special spotlight on those (older, more experienced, senior) managers that firmly rely on their experience in purchasing. In this type of task they are strikingly outperformed by their young professional colleagues. Moreover, in the I&L-1 experiment we find that managers that rely on their first impressions and on their gut feeling perform less well than those who are more reluctant to do that.

The research finding experienced that experts do not perform better in professional judgments, predictions and decisions than those with less experience, does not only hold in purchasing alone. There are a large number of domains for which the literature has shown similar results. Based on a meta-analysis of 55 research projects on judging the performance of professionals in the field of psychology and psychiatry, Garb (1989, 1998) comes to the conclusion that experienced professional psychologists may perform somewhat better than super lay persons, such as freshmen students or secretaries, but that their performance is hardly ever better than that of graduate students who received only a moderate level of training. In general, a little bit of training helps, but gaining a lot of experience after that initial stage hardly ever does. Similar conclusions have been reached in the field of medicare (Gustafson, 1963; Kundel and LaFollette, 1972; Shortliffe et al., 1979), as well as in many other domains of expertise (Camerer and Johnson, 1991). Characteristic for the field of purchasing is that we reach even sharper conclusions: in making professional judgments many experienced professional experts perform extremely poorly. How can these findings be explained? Again, we search for answers in the literature on expertise.

One feature that is characteristic for the field of purchasing management, and for many other fields of professional expertise, is the lack of frequent and direct feedback. Without systematic feedback it is hard, and sometimes even impossible to learn from one's own mistakes and misjudgments, no matter how experienced one is. Moreover, following Camerer and Johnson (1991), characteristic for experienced professional experts in such a context without immediate and accurate feedback is that, compared to lay persons, novices and young professionals

- experts are more selective in their search for information,
- experts store information much faster,
- experts have a more active pattern of contingent search: Subsets of variables are considered in each case, in different sequences,
- experts use less information, and the information used is more combined in a non-linear way; they often use configural choice rules,

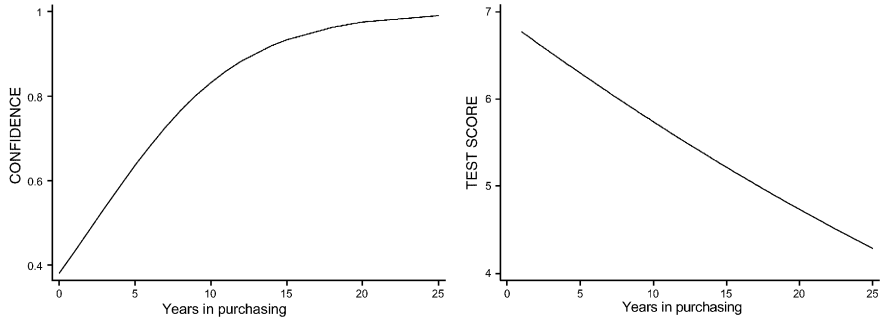

Fig. 2. Confidence in the accuracy of judgments (left) and score on the test (right) by years in purchasing (best fit cubic regression line).

- experts compare the information given with their own knowledge base of prototype cases; their choices are often characterized by an over-generalization of specific prototype examples,
- experts use more "broken-leg cues" (I.e., cues of the kind that, if encountered, ensure that the expert immediately knows—or thinks to know—that other cues are now irrelevant. For instance, to predict if I will run the marathon tomorrow, a "broken-leg cue" would be that I have a broken leg.).

This type of decision-making may be adequate in highly specific contexts, for instance in turbulent market situations and situations that are characterized by abrupt change, but in many other contexts this type of information processing will be less profitable. Decision-making in purchasing is too "noisy" for professionals to acquire expertise and use the abovementioned decision strategies with success. Generally, decisions in purchasing are made in a context where feedback is lacking, where it is not really clear which case characteristics are good predictors, where measurement of what could be the relevant case characteristics is often lacking, and where the outcome is not strictly deterministic but probabilistic instead. In such a setting, it is not that surprising that cold, calculating formulas do a better job.

It may seem as if the judgments by purchasing professionals are altogether useless when compared to a reasonably selected formula. As we will show, this is not the case. When carefully employed, the professional's judgments can be used to improve purchasing decisions.

5. Improving experts

Another way to understand the behavior of professionals in purchasing, better, is by trying to find out how to improve upon their scores. There are several ways of doing this (Kleinmuntz, 1990), but at the heart of most of these lie two ideas. The first one is the idea that experts, just like humans in general, are "too volatile" in their judgments. This can be understood as follows. The formula, which is a linear prediction, provides an appropriate "average judgment" across cases. Even if the formula performs very well, you can be sure that if you judge like the formula, you will certainly make mistakes in many instances: fitting a straight line through a cloud of dots means missing most of the dots, even if the line is on average close to many of them. Humans are reluctant to judge in what they feel is a "too rigid" way, because they strive (too hard for their own good) to give the proper judgment each and every time. Unfortunately, this makes them perform worse on average in the long run (cf. Kleinmuntz, 1990; Camerer and Johnson, 1991). This is the basis of the first way to try to improve human experts. On can try to somehow partial out the excessive variance of human experts. The second idea is that, although experts generally deviate from linear predictions in too strong a manner, there could be some merit in combining the judgments of a model with those from experts to get a model-plus-expert prediction that outperforms both the model and the experts. This might work because in this combination the model takes care of an overall linear fit that is reasonable, whereas the experts' intuition could capture the deviations from the linear model (Blattberg and Hoch, 1990; Goodwin, 2002; Morris, 1986). Of course, whether these ideas, reducing variance and combining model and expert work here, is an open question. In other fields, such variants have shown to be a useful improvement (e.g., Ashton and Ashton, 1985; Clemen, 1986; Edwards, 1962; Einhorn, 1972; Makridakis and Winkler, 1983; Showers and Chakrin, 1981). We now try several combinations of these ideas to test if they improve expert forecasting.

5.1. Creating an 'expert system' by averaging across experts

One way to try to improve upon the predictions of experts through reducing variance, is by using the estimates of the whole group of experts to calculate the 'average expert estimate' as dependent on the characteristics of the cases, and then using this 'average expert model' to predict the likelihood of problems in the different cases (Camerer, 1981; Hill, 1982; Hogarth, 1978). This represents on a small scale how one could conceive of creating a knowledge base by combining expert judgments. In a way, this expert model can be seen as the combined knowledge of all participating experts. Across all five experiments, this procedure leads to a model with a Spearman coefficient of 0.23, only a small and borderline statistically significant ($p = 0.08$, one-sided) improvement over using the estimates of the separate experts. Nevertheless, the formula still outperforms the expert system, with a score of 0.32.

5.2. Creating an 'expert system' by averaging within experts across cases

A comparable expert model can be made in a slightly different way. *For each expert*, we try to calculate a model based on his or her estimates on the eight cases. Obviously, given only eight cases per expert, it makes no sense to try to estimate a model using 14 different variables (which is the number of case characteristics we have). We, therefore, restrict ourselves to three of the stronger predictors of eventual problems: the volume of the transaction, the duration of the past relationship, and the ability to judge the relation between price and

Table 6
Spearman correlations between predictions on the degree of problems and the actual degree for different models using the expert judgments (higher is better). Results are averaged across the five experiments and for purchasing professionals only

Experts	Formula	Expert system 1	Expert system 2	Formula + experts	Formula + exp. system 1
0.19	0.32	0.23	0.19	0.27	0.36

quality. If we use these three variables to predict a best fit for each expert, and subsequently use this model's predictions instead of the expert's own predictions, we end up with a Spearman correlation of 0.19. This is equally accurate as the experts' own predictions were. In other words, an 'average expert model' with weights on only three (aptly chosen) variables performs just as well (or just as bad, if you prefer) as the expert. The formula's score, 0.32, is still out of reach. The latter comparison is not really fair, because the formula uses more than these three variables. If we confine the formula to the same three variables, it's score is 0.25, still better than the expert.

5.3. Combining experts' judgments with the formula

Next, we try the second possible improvement, by blending the formula's predictions with the predictions of experts. In accordance with Blattberg and Hoch (1990) we first calculate a new set of predictions for each case, by taking the average of the formula's prediction and the prediction of the expert. Now the Spearman correlation increases to 0.27, a substantial improvement over the experts' own judgments, but still lower than the formula's 0.32. There is, however, an interesting finding here. If we use weights other than 50% model and 50% expert, it is indeed possible to get an improvement that even outperforms, though only slightly, the formula. A weighting of about 90% model with 10% expert yields a Spearman correlation of 0.35. This is not a fair comparison, since the 90–10 division was found in a completely 'data-driven' way. Nevertheless, it does show that it may be possible to use the experts' judgments to improve upon the formula per se. This promising finding leads us to try a different and less data-driven way of combining formula and expert.

5.4. Combining an 'expert system' with the formula

Since the optimal weighting when using a combination of the formula and the expert judgments was 90–10, perhaps the expert judgments by themselves are still too volatile. But what if we combine the formula with the 'expert system' mentioned in Section 5.1. It is conceivable that in this way we can get the best of both worlds. The volatility in expert judgments is decreased, while maintaining enough of the expert intuition that caused the favorable results in the previous attempt. Thus, we first calculate—as we did above—a formula that represents the 'average expert' by regressing the experts' judgments on the case characteristics. Then, we combine this average expert judgment with the formula, with equal weights for both. It turns out that this indeed gives even better performance: the Spearman correlation now equals 0.36, a small but statistically significant increase ($p = 0.024$, onesided) that is stable across the five different experiments.

A summary of the results is presented in Table 6.[3]

Note that the original formula performs best, except for one alternative model, i.e. the formula combined with the expert system, the last of the abovementioned alternative models. The difference is hardly substantial, but at least we do find some evidence here that the decisions of experts can be incorporated in the overall prediction to some effect.

6. Conclusion and discussion

The finding that a formula predicts the prevalence of problems better than purchasing professionals has been shown to be a robust finding across different experiments. Our analyses and review of part of the relevant literature also give some feel for the causes underlying the relatively poor performance of professionals. Experts process a judgment task differently than do non-professionals, tend to relate their judgments to selectively available cases, and use larger chunks of information at the same time. Even though they do often not outperform less experienced colleagues, they are generally more certain about the accuracy of their outcomes. When looking closer at the data, we even find some evidence that over the course of their career and with increasing experience, purchasing professionals develop habits that hamper rather than facilitate providing accurate predictions. An accurate overall prediction typically involves consistently weighing the available case characteristics, and professionals are less inclined to do that as their experience progresses.

[3]Again, one could argue that the Spearman (rank) correlation is an unnecessary simplification here, since the prediction criterion has an interval scale. The results reported do not depend on whether one uses rank correlation or Pearson correlation.

One might be tempted to think that at this point one cannot but conclude that the judgments of purchasing professionals are superfluous altogether. And undeniably, most potential improvements of the judgments of professionals do *not* lead to judgments that outperform the cross-validated formula. However, we do find evidence that a sensible combination of both the formula and the expert judgments may improve upon the professional and may even outperform the formula. This involves first combining the expert judgments into an 'average expert', and then combining this with the formula.

Taken together, our analyses show that purchasing professionals are not an exception to the general rule in the scientific literature that computer models outperform professionals when it comes to the kind of clear cut decision tasks as described here. Though this may seem to paint a bleak picture of the abilities of purchasing professionals, one should not stretch our findings beyond their limits. Let us first stress that we are not claiming that—in general—computer models are better purchasing professionals. There is much more to being a purchasing professional than just making single shot decisions of the kind we discussed, and certainly many of the challenges purchasing professionals have to deal with are beyond the realm of a computer model. Our conclusions are confined to tasks of the kind as in our experiment. These are tasks where the decision is clear-cut (e.g., Should I use tenders and how many?), and where there is at least some data on past performance available for a model to be constructed. We discuss some of these issues in the section on practical implications below. Several factors typically make such decision tasks more difficult for purchasing professionals as opposed to computer models: they work in an environment where immediate and precise feedback is generally lacking, and where many factors influence the eventual outcome so that the optimality of their own behavior cannot be clearly distinguished.

Several counterarguments to our results can be thought of. First of all, in our experiment the professionals are forced to make their decisions in a context that is certainly more abstract than they are used to. Although our results still hold if we only use the data of the professionals who claimed to be certain about their answers, it is possible that their judgments would improve and perhaps surpass the judgments of the model with increasing information about the transaction. We aim to test this in a future experiment. A second counterargument we can think of is that in our experiment there are no real incentives to make the right decisions. Purchasing professionals are used to decide in situations with a lot of money at stake, and are perhaps less inclined to think carefully when all there is to gain is a high test score. It is difficult to argue against this given available data. Closer inspection of the cases where professionals took the test more than once shows that their results do not tend to improve on second runs. This could perhaps be interpreted as a sign that extra attention does not help much, but we admit that this is not very compelling. Again, this is something to take into account in a future experiment.

7. Implications for practitioners

Our findings corroborate the idea that it would be wise to at least take a closer look at many of the decisions and predictions in purchasing, or even management decisions in general. There is a definite gain in using decision support of a specific kind: devising a formula that systematically combines relevant inputs. A simple formula will often outperform the purchasing expert's intuition and experience. As we have experienced first hand, many purchasing professionals are surprised about the improvements in accuracy that can be accomplished by letting a formula decide, especially for the decisions that they feel are typically their area of expertise. In particular, one should focus on decisions and predictions where, in practice, immediate feedback of the correctness of decisions is lacking or imprecise. In these cases, it is difficult, if not impossible for professionals to learn from experience, and professionals will have a hard time building up expertise, even though they may feel otherwise, and even though others may judge them as experts. Typically, for many situations in purchasing and managerial decision-making, one can never know what would have happened in case a different decision had been made. This implies that in fact most of the predictions and decisions in purchasing and management lack this immediate and adequate feedback that is necessary to build up expertise, and are a serious candidate for computerized decision-making.

Of course, when data on which to base a prediction or decision are not available, one can only use judgments by professionals. Note, however, that in these cases it is generally possible to improve upon the separate expert judgments by aggregation of expert judgments to reduce excessive variance. What remains throughout all our experiments and others in the literature, is that a major improvement in purchasing decisions can be made if at least some (representative) data that are relevant to the decision are available (and used for prediction). A rather straightforward formula based on data from past cases soon outperforms professional judgment.

The kinds of decisions for which one could, or perhaps should, consider making use of formulas instead of professional judgment are typically decisions that occur rather frequently (otherwise it is less feasible to gather the data from previous cases). For instance, we feel an improvement could be made for questions such as the choice for the optimal number of tenders, the

extent to which one invests in the management of a purchasing transaction, which of a set of purchasing transactions to invest more time in, and so on. Decisions that are 'one of a kind' (Should we follow through with this merger? Should we become an organization that operates internationally?) are less suitable for such an approach, though principally one could conceive of formula's being devised on a large set of these kinds of decisions from other organizations. For those willing to invest an extra effort in improving decision-making even more, combining the formula with an 'agent formula' that mimics the average behavior of experts, there is a (relatively small) extra gain in accuracy.

A final implication relates to the education and training of professionals. We feel that the first step to improvement is awareness of the problem. To start with, it would therefore make sense that the education and training of purchasing professionals at least includes the message that the judgments and decisions of purchasing professionals are not as good as they can be, and can be improved upon by using computer models. In practice this will be hard to get across, let alone implement. This is because this message hits professionals in an area in which they consider themselves the experts, which makes it difficult to accept. Moreover, the reasons for the relative inferiority in judgments are so ingrained in humans that they are hard to overcome. Given our result that experience does not improve judgments, we think the most progress can be made by educating young professionals to be flexible and open-minded about using computer models to their advantage.

Acknowledgements

Snijders gratefully acknowledges support by the Royal Netherlands Academy of Arts and Sciences (KNAW).

Appendix A. Vignette construction and measurement

The vignettes were taken from our database of Purchasing Transactions of the External Management of Automation, a large scale survey on the purchase of IT-products by Dutch SMEs (5–200 employees). The sampling frame was a business-to-business database of Dutch SMEs that contained information about the characteristics of these SMEs with respect to automation. The database can be considered to be representative for the Dutch population of SMEs (see Batenburg, 1997). Care was taken to achieve high response rates: firms were contacted by phone and if a respondent agreed to fill out a survey on a specific purchasing transaction, field workers delivered the survey on the agreed upon date and were instructed to leave with the filled out survey. If the respondent was willing to fill out a second survey on a different purchasing transaction, the field worker was to leave a questionnaire and a response envelope so that the respondent could fill out this second case at his or her convenience. Eventually, the average response rate to the telephone interview was 67% (902 out of 1,335). Multiplied with the field response rate of 87% (788 out of 902), the total response rate equaled 59% (788 out of 1,335). This is a high response rate in comparison with other surveys among organizations (cf. Kalleberg et al., 1996: chaps. 1–2). Non-response analysis showed that the response group is not biased on crucial firm characteristics such as size, industry or region. The codebook, which includes the questionnaires, is downloadable from http://www.fss.uu.nl/soc/iscore.

Per transaction, over 300 issues were measured, including the ones mentioned in the vignettes, such as the number of tenders, the number of days involved in negotiating and contracting, etc. The problems associated with the transactions were measured by having respondents check on a 5-point scale the (degree of) problems that had occurred. Based on results in the pilot phase the respondents could choose from "late delivery", "over budget", "product incomplete", "product too slow or too confined", "specification not as agreed", "incompatible with other systems", "sloppy installation", "after-sales slow or missing", "service slow or missing", "necessary adjustments slow or missing", and "insufficient documentation". Our empirical analysis turned out that there is a single dimension underlying these problems—a principal component analysis yielded a clear single component. Our data show that, on average, a large number of problems and a high degree of problems go together. The degree of "problematicness" is therefore measured as the average score on this list of problems. We then classified all cases according to their score into eight categories of similar size, ranging from 1 = not very problematic to 8 = very problematic. Each individual received one (randomly chosen) vignette from each category, in a random order.

References

Ashton, A.H., Ashton, R.H., 1985. Aggregating subjective forecasts: some empirical results. Management Science 31, 1499–1508.

Batenburg, R., 1996. The external management of automation 1995. Codebook of MAT95. Utrecht University. ISCORE Paper No. 58, 218pp.

Batenburg, R.S., 1997. The External Management of Automation. Codebook of MAT95. ISCORE Paper No. 58. Department of Sociology, Utrecht University.

Batenburg, R., Raub, W., Snijders, C., 2000. Contacts and contracts: temporal embeddedness and the contractual behavior of firms. Utrecht University. ISCORE Paper No. 107, 34pp. http://www.fss.uu.nl/soc/iscore.

Blattberg, R.C., Hoch, S.J., 1990. Database models and managerial intuition: 50% model + 50% manager. Management Science 36 (8), 887–899.

Buskens, V., Batenburg, R., 2000. The external management of automation. Codebook for the combined data from The Netherlands and Germany. Utrecht University. ISCORE Paper No. 175, 258pp. http://www.fss.uu.nl/soc/iscore.

Camerer, C., 1981. General conditions for the success of bootstrapping models. Organizational Behavior and Human Performance 27, 411–422.

Camerer, C., Johnson, E., 1991. The process-performance paradox in expert judgment. How can experts know so much and predict so badly? In: Ericsson, K.A., Smith, J. (Eds.), Towards a General Theory of Expertise: Prospects and Limits. Cambridge University Press, Cambridge.

Clemen, R.T., 1986. Calibration and the aggregation of probabilities. Management Science 32, 312–314.

Dawes, R.M., 1971. A case study of graduate admissions: application of three principles of human decision making. American Psychologist 26, 180–188.

Dawes, R.M., 1979. The robust beauty of improper linear models in decision making. American Psychologist 34, 571–582.

Dawes, R.M., Faust, D., Meehl, P.E., 1993. Statistical prediction versus clinical prediction: improving what works. In: Keren, G., Lewis, C. (Eds.), A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues. Lawrence Erlbaum, Hillsdale, NJ, pp. 351–367.

Edwards, W., 1962. Dynamic decision theory and probabilistic information processing. In: Kleinmutz, B. (Ed.), Formal Representation of Human Judgment. Wiley, New York, pp. 17–52.

Einhorn, H.J., 1972. Expert measurement and mechanical combination. Organizational Behavior and Human Performance 7, 86–106.

Garb, H.N., 1989. Clinical judgment, clinical training, and professional experience. Psychological Bulletin 105, 387–396.

Garb, H.N., 1998. Studying the Clinician. Judgment Research and Psychological Assessment. American Psychological Association, Washington.

Glaser, R., Chi, M.T.H., 1988. Overview. In: Chi, M.T.H., Glaser, R., Farr, M.J. (Eds.), The Nature of Expertise. Lawrence Erlbaum Associates Inc., Hillsdale, N.J., pp. xv–xxviii.

Goodwin, P., 2002. Integrating management judgment and statistical methods to improve short-term forecasts. Omega: The International Journal of Management Science 30, 127–135.

Grove, W.M., Meehl, P.E., 1996. Comparative efficiency of informal (subjective, impresionistic) and formal (mechanical, algorithmic) prediction procedures: the clinical-statistical controversy. Psychology, Public Policy, and Law 2 (2), 293–323.

Grove, W.M., Zald, D.H., Lebow, B.S., Snitz, B.E., Nelson, C., 2000. Clinical versus mechanical prediction: a meta-analysis. Psychological Assessment 12 (1), 19–30.

Gustafson, J.E., 1963. The computer for use in private practice. Proceedings of the Fifth IBM medical symposium. IBM Technical Publication Division, White Plains, NY, pp. 101–111.

Hill, G.W., 1982. Group versus individual performance: are N + 1 heads better than one? Psychological Bulletin 91, 517–539.

Hogarth, R.M., 1978. A note on aggregating opinions. Organizational Behavior and Human Performance 21, 40–46.

Kalleberg, A.L., Knoke, D., Marsden, P.V., Spaeth, J., 1996. Organisations in America. Analyzing their Structures and Human Resource Practices. Sage, London.

Kleinmuntz, B., 1990. Why we still use our heads instead of formulas: toward an integrative approach. Psychological Bulletin 107 (3), 296–310.

Kundel, H.L., LaFollette, P.S., 1972. Visual patterns and experience with radiological images. Radiology 103, 523–528.

Makridakis, S., Winkler, R.L., 1983. Averages of forecasts: some empirical results. Management Science 29 (9), 987–996.

Meehl, P., 1954. Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence. University of Minnesota Press, Minneapolis.

Meehl, P., 1986. Causes and effects of my disturbing little book. Journal of Personality Assessment 50, 370–375.

Morris, P.A., 1986. Observations on expert aggregation. Management Science 29, 24–32.

Rooks, G., 2002. Contract en conflict. Het strategisch management van inkooptransacties. (Contract and conflict. The strategic management of purchasing transactions.) Amsterdam: Thela Thesis.

Shanteau, J., 1987. Psychological characteristics of expert decision makers. In: Mumpower, J.L., Renn, O., Phillips, L.D., Uppuluri, V.R.R. (Eds.), Expert Judgment and Expert Systems. Springer, Berlin.

Shanteau, J., Peters, J.M., 1989. The 3 C's of Expert Audit Judgment: creativity, confidence, and communication. In: Mock, T. (Ed.), Paper presented at the February 1989 USC Audit Judgment Symposium. University of Southern California School of Accountancy.

Shortliffe, E.H., Buchanan, B.G., Feigenbaum, E.A., 1979. Knowledge engineering for medical decision making: a review of computer-based decision aids. Proceedings of the IEEE 67, 1207–1224.

Showers, J.L., Chakrin, L.M., 1981. Reduceing uncollectible revenue from residential telephone customers. Interfaces 11 (6).

Snijders, C., 2003. Computer verslaat inkopers [Computer beats purchasers] Tijdschrift voor Inkoop & Logistiek. Journal of Purchasing and Logistics 19 (6), 47–48.

Snijders, C., Tazelaar, F., Batenburg, R., 2003. Electronic decision support for procurement management: evidence on whether computers can make better procurement decisions. Journal of Purchasing and Supply Management 9, 191–198.