



**Aviation Human Factors Division
Institute of Aviation**

**University of Illinois
at Urbana-Champaign
1 Airport Road
Savoy, Illinois 61874**

**Is There a Magic Number 7
(to the Minus 1)? The Benefits of
Imperfect Diagnostic Automation:
A Synthesis of the Literature**

**Christopher D. Wickens
and Stephen R. Dixon**

**Technical Report
AHFD-05-01/MAAD-05-1**

September 2005

Prepared for

**Micro Analysis and Design
Boulder CO**

Contract ARMY MAD 6021.000-01

ABSTRACT

This review of the literature examines in a quantitative fashion, how the level of imperfection or unreliability of diagnostic automation affects the performance of the human operator who is jointly consulting that automation and the raw data itself. The data from 20 different studies were used to generate 35 different data points that compared performance with varying levels of unreliability, with that of a non-automated baseline condition. A regression analysis of benefits/costs relative to baseline was carried out, and revealed a strong linear function of benefits with reliability. The analysis revealed that a reliability of 0.70 was the “crossover point” below which unreliable automation was worse than no automation at all. The analysis also revealed that performance was more strongly affected by reliability in high workload conditions, implicating the role of workload-imposed automation dependence in producing this relationship, and suggesting that humans tend to protect performance of concurrent tasks from imperfection of diagnostic automation.

INTRODUCTION

It is well known that automation can serve as an aide to human performance (e.g., Sheridan 2002). In the current article, we consider two interacting factors that can moderate this automation benefit to human performance, both intuitively obvious on the surface, but more complex in terms of their quantitative effects.

The first factor is the level of “**workload**” experienced by the human operator. Here we define workload, as the load imposed on limited resources of the unaided (without automation) human operator, in what we describe as the “baseline” or “manual” condition. This load can be imposed from two qualitatively distinct sources: the single task **difficulty** of the task that might otherwise be automated, and the **concurrent task load** in which the baseline (vs. automated) task is performed. In both of these two cases, the automation benefits are likely to increase, to the extent that the single task to be automated is more difficult (Maltz and Shinar 2003, Dixon and Wickens 2004a), and/or that concurrent task load is imposed (Parasuraman et al. 1993).

The second factor is automation **reliability**. There is little doubt that total human-system performance will be quite good if automation is perfect. Conversely, when carrying out a difficult task, performance will be poor when automation is so unreliable as to be useless, and may actually harm performance relative to a purely manual baseline condition. However in between these extremes lies a range of reliability levels, where the benefits of automation over the baseline may be uncertain, and it is this range that is the focus of the current review.

Thus in the current paper, we consider the combined influence of reliability and workload on joint automation-human performance. The issue is more generally relevant to a wide range of applications, simply because the joint influence of these variables has rarely been addressed in a quantitative fashion that might be amenable to computational modeling.

The importance of computational modeling is highlighted by efforts to predict acceptable levels of human-system performance as the human is asked to do more, but be assisted by automation. A specific example of this is in the control of multiple uninhabited air vehicles (UAVs; Wickens and Dixon, in press, Dixon and Wickens in press). How many is too much? If

automation were perfectly reliable, it would be hard to specify an upper limit. But known imperfections in automation render this a meaningful question to ask. While time consuming pilot-in-the-loop simulation can provide some answers, models of human response to imperfect automation can complement these answers. And a valuable parameter in such a model would be a “threshold” of reliability below which automation benefits disappear.

Of course such a “reliability threshold” is unlikely to be exact. Numerous contextual factors such as human operator knowledge and payoffs may well moderate this value considerably (Lee and See, 2004). These factors vary across the studies that are synthesized in the current review, and for the most part are not systematically quantified here. Instead, they are treated as “noise”, which contributes to variance in the observed benefits or costs. However one contextual variable in particular is explicitly extracted and accounted for; the level of **task load** during which the automation is implemented.

Automation, of course, covers a vast range of functionalities, including autopilots, calculators, expert systems, etc. In the current writing, we narrow our scope in this literature synthesis to one common generic type of automation: **diagnostic aiding**. Within the four stage taxonomy of automation proposed by Parasuraman et al. (2000), diagnostic aiding refers both to stage 1 automation (e.g., filtering or focusing attention on information deemed to be of interest), and stage 2 automation (e.g., forming inferences of the state of the world, by integrating information). It excludes stage 3 automation, (recommending or selecting a course of action), and stage 4 automation (implementing that action). Thus the common element in the automation that we examine here is that such automation categorizes environmental elements into one of two states, which we can generically label as “target” and “non-target” states (Swets and Pickett 1982).

Such diagnostic systems possess several features that render them desirable targets for the current modeling research:

- Unlike stage 3 automation (choice), performance does not depend upon values or utilities, which may be highly subjective in nature, and hence, hard to quantify.
- Unlike stage 4 automation (action execution), automation benefits can be represented purely as cognitive benefits, and do not involve the motor system, and issues of physical debilitation and fatigue.
- Diagnostic automation often is asked to perform tasks for which unreliability and imperfection is **expected** (even if it is not desirable); hence reference to “unreliability” does not signal a catastrophic failure, in the same manner as a failed autopilot; rather, automated diagnosis may be challenged by such inevitable factors as predicting an uncertain future state (Thomas Wickens and Rantanen 2003), or integrating probabilistic diagnostic cues. In both cases, a small number of imperfections that are either automation “misses” (a critical signal or event is not detected by automation) or “false alerts” (alert in the absence of an event), are both tolerated and expected. We ask here how large this number can be before automation loses its effectiveness.

- Reliability can be objectively quantified. In diagnostic systems, this quantification is often done in terms of signal detection theory parameter “d-prime” (Swets and Pickett, 1983; Swets, 1995). However here our focus is on simply a proportion correct measure that varies between 0 and 1, an advantage because of the ready association of this scale with the scale typically used in reliability engineering (e.g., # correct operations/total number of operations, or 1 – failure rate)
- Such diagnostic systems are quite ubiquitous in a variety of applications, related to systems such as warning devices (Stanton, 1993), medical diagnostics (Swets and Pickett, 1983), intelligent target aiding (Maltz and Shinar, 2003, Wiegmann and Madhavan, 200X) collision alerts (Xu Rantanen and Wickens, in press) and statistical tests (Keppel and Wickens, 2005).

In conclusion, the goal of this paper is to establish the extent to which *there is a value of diagnostic reliability below which automation becomes useless, or even worse than baseline performance* and the extent to which this value may be modulated by one important contextual variable: task workload. The search for such a reliability value is guided by the knowledge that constants like this, can be of considerable value in developing computational models, such as, here, a computational model of human-automation performance. As well, such constants can serve design guidelines, just as the “magic number 7” constant representing the limits of working memory has, for half a century, guided designers on ways to avoid working memory overloads (Miller 1956). For example, it may serve as guidance for system managers to accept or reject the acquisition of such a system. Rather than presenting a narrative text review of the various studies in the literature, we instead integrate this knowledge in a more quantitative form, in the manner described in the following section.

PROCEDURE

We initially sought all studies we could locate that examined imperfect automation. In further filtering the literature that was revealed in this search, we established several criteria. In particular, we ultimately included only studies in which:

1. Stage 1 or stage 2 automation was examined in a way that allowed the output to be expressed in a signal detection matrix, so that the proportion correct measure could be derived (e.g., $1 - [FA + M]/[Total\ events]$). This allows quantification of the level of automation performance on a single metric.
2. The paradigm was one in which the “raw data” were available for perception by the human operator, in parallel with the output of the diagnostic aide, as described by Sorkin and Woods (1985) (although the extent to which the operator chose to perceive those raw data was not an issue that guided our filtering of relevant literature).
3. Two or more conditions were assessed, one of which needed to be an imperfect reliability condition ($r < 1.0$), and the second of which was a **baseline, non-automated** condition, in which the operator examined the raw data only. This allows assessments of the benefits (or costs) of imperfect automation. We also considered studies in which two or more levels of imperfection were examined, as well as those in which a condition of

perfect automation was included, even if these studies did not include the most critical baseline.

4. The operator using the imperfect automation system was allowed an opportunity to appreciate that the imperfections existed, either through training and exposure to automation errors, or through prior instructions.

It is noteworthy that one very important class of imperfect diagnostic automation studies are excluded from this review by virtue of criterion 4, and those are studies that have looked exclusively at automation **complacency**, as reflected in what we call “first failure effects” (Wickens 2000). These are circumstances in which the operator encounters perfect automation for some period of time, and then “complacency” or overtrust in automation is reflected by the very poor response to the automation when it initially fails (Molloy and Parasuraman 1996, Metzger and Parasuraman 2005). While this is a very important phenomenon in its own right it does not reflect the “steady state behavior” of the operator who is **calibrated** to the level of imperfection of the automation (Wickens Gempfer and Morphew 2000), a state that should be the goal of most training programs designed for automation with a particular system.

5. Finally, a fifth important attribute (although not one used as a basis for study filtering) was the existence of a manipulation of workload, through single task difficulty, or concurrent task load. In the latter case, it was important for us to be able to assess the benefits to concurrent task performance of imperfect automation, as the latter, presumably, might influence the attention demands of the automated task.

Thus, while a large number of the original tally of over 40 studies of imperfect automation were filtered out by these criteria, we believe that the current subset represents a sufficiently important class, and that their results are sufficiently consistent (reflected by the restricted standard error) as to be meaningful and useful. We certainly invite other investigators to extend our methodology with regard to stages 3 and 4 automation.

RESULTS

A total of 22 studies were identified that satisfied the criteria listed above. These studies are listed alphabetically in table 1, and full references are provided in the reference list. Each study is also assigned a numeric code, used in later analysis and representation.

Table 1 presents data in the following manner. In column 1 is the study and its ID number. Note that sometimes a given study generates two entries (and separate ID numbers) when two different levels of task workload are imposed within the study. The space to the right indicates an ordinal scale of joint human-system performance. It contains a “**B**” indicating the level of baseline performance for a given study. Automation conditions are indicated by a value corresponding to the reliability of the automation. An entry to the right of the baseline implies that that reliability condition produced better performance; one to the left indicates worse performance. A value directly below indicates equivalent performance. The ordinal relations are reinforced by a “>” or “<” symbol. Sometimes the performance values, used to generate the ordering were not directly provided in the article, but required derivation by the current authors. (For example sometimes the net automation-human system performance was reported separately

Table 1: Summary of studies and their automation reliability levels.

	Performance Effects	Concurrent Tasks
1. Bliss	.50 < .75 < 1.0	
2. Davison	B < .70	.70 < B
3. Dingus		B (RT)
4. Dixon(low) ^a	.64 < B (RT) .80 .64 < B (AC) .80	.64 = .80 B < .80(AC) .64
5. Dixon(high) ^b	.64 < B (RT) .80 .64 = .80 < B (AC)	.64 < .80 < B (RT)
6. Dixon05	.60 < B (RT) .60 < B (AC) B .60(RT) B .60(AC)	.60 < B B .60
7. Dzindolet	.65 = .75 = .90	
8. Galster(low)	B .67	
9. Galster(high)	B < .67	
10. Lehto	B < .70 < .90	
11. Maltz & Meyer	B < .91 < 1.0(inferred) .75	
12. Maltz03(low)	.60 < .80 < .90 < B	
13. Maltz03(high)	.60 < B < .90	
14. Maltz04 ^c	Low < High	
15. Meyer	.67 < 1.0	

	Performance Effects	Concurrent Tasks
16. Para(93) ^d	.56< B <.87	.56=.87
17. Skitka	B <.88	B .88
18. St. John	B <.88	
19. Wickens & Dixon(low) ^e	B <.60(RT) B (AC) .60 B <.60(RT) B (AC) .60	B (RT) .60 .60< B (AC) .60< B (RT) .60< B (AC)
20. Wickens & Dixon(high) ^e	.60< B (RT) B (AC) .60< B (RT) .60< B (AC)	B (RT) .60 B (AC) .60 B (RT) .60 B (AC) .60
21. Wiegmann	B <.86	
22. Xu	B ^f <.83	
23. Yaccov	B <.80=.95	
24. Yeh	B .70	

^aResults of two experiments are combined with $r = .67$ and $r = .60$. Baseline performance was only measured in one of these.

^bConcurrent task accuracy only measured low workload.

^cPrecise value cannot be calculated because only number of false alarms given. No base rate.

^dBaseline computed from Molloy 96. No baseline available for concurrent task.

^eThe first and second pair of RT/accuracy readings represent between subject-conditions in which .60r automation was biased to produce false alarms and misses, respectively.

^fFor the half of the participants who depended upon automation.

for trials in which automation erred and in which automation was correct and it was necessary for us to compute the net system performance as a weighted mean of performance on the two types of trials, weighted by the relative frequency of correct vs. erroneous automation). Furthermore, sometimes the actual reliabilities themselves were not reported in a proportion format, but required conversion from a “d prime” number.

Finally, on the far right, where relevant, is presented any information available about concurrent task performance, using the same representation as used for the automated task.

The data in table 1 have been re-plotted in graphical form in figure 1a. Arrayed along the X axis of figure 1 is the reliability of imperfect automation. On the Y axis, we have plotted the cost (below the X axis) or benefit (above) of imperfect automation relative to the baseline, whose performance (proportion correct) is indicated on the axis itself. Generally only three Y axis values are employed, +1 for a benefit, 0 for equivalence, and -1 for a cost. Cost and benefit were derived from statistical significance (0.05 criterion), as reported by the authors. We did not apply further estimates of effect size such as would be done in a meta-analysis (Rosenthal, 1991). When two or more imperfect conditions differ significantly between themselves (in cost or benefit), and both lie on the same side of the baseline axis, then we assign a 1 to the closer value (to baseline) and a 2 to the further value, assuming that the latter has a greater benefit (or cost) than the former.

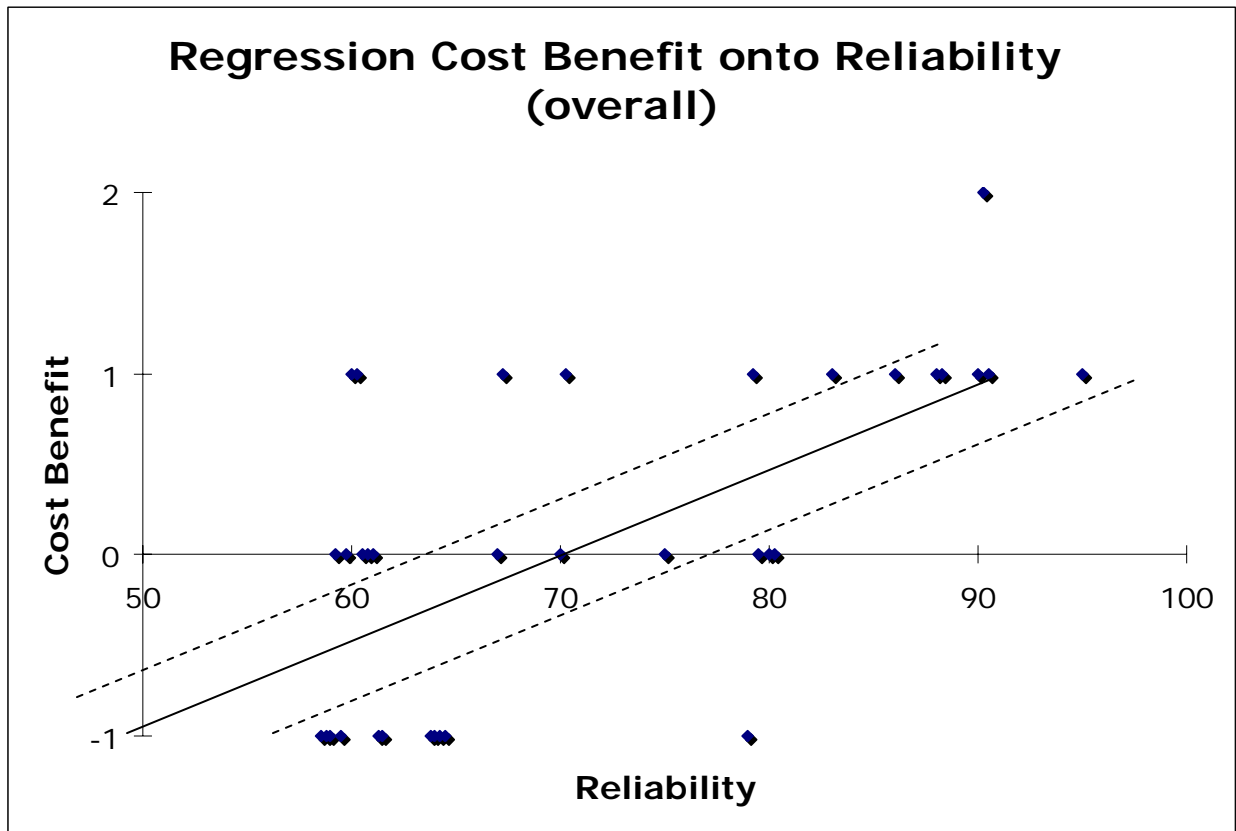


Figure 1a. Regression of benefits/costs relative to baseline, on automation reliability with a 95% confidence interval.

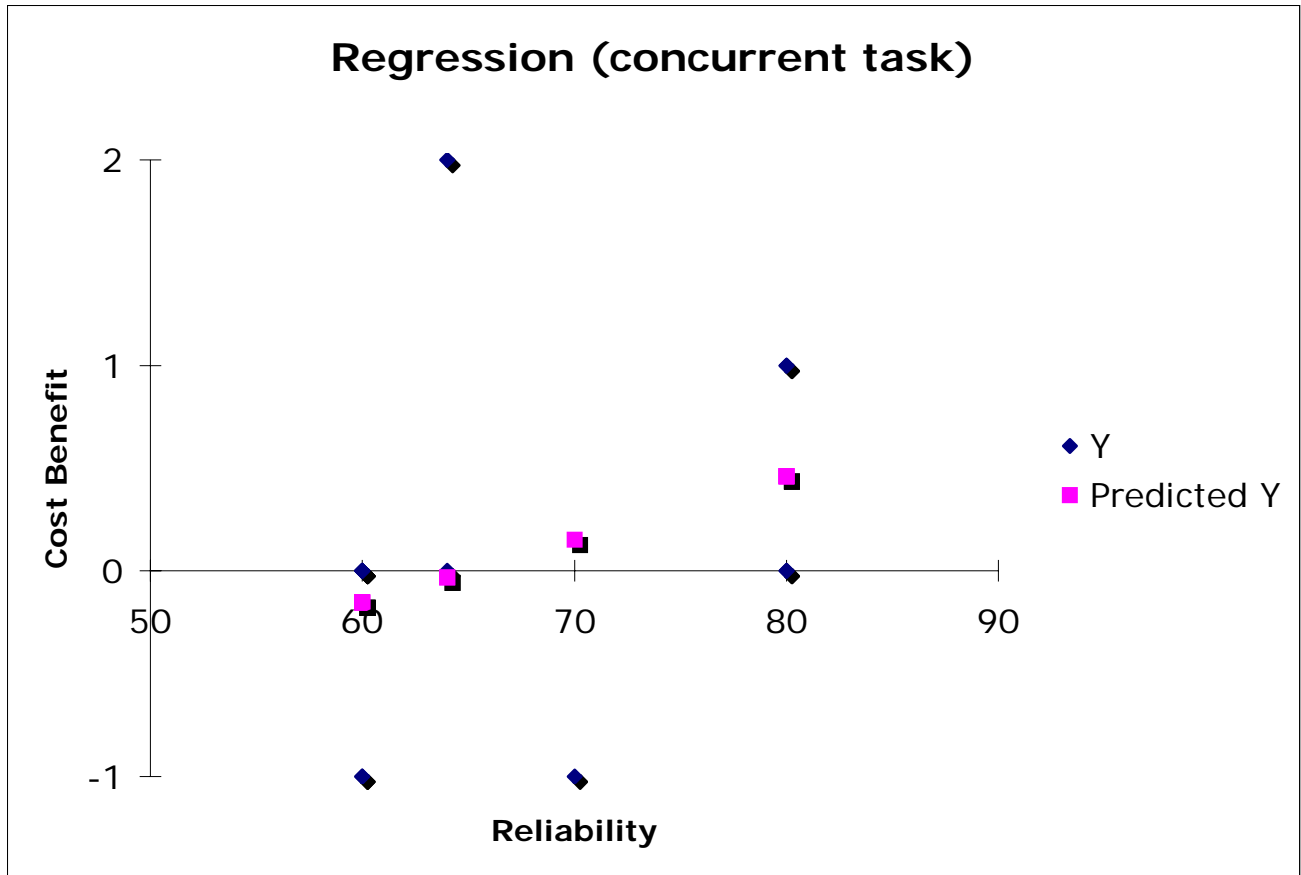


Figure 1b. Regression analysis of concurrent task performance.

Not all studies listed in table 1 are plotted in figure 1. We did not plot data from those studies that did not include a baseline condition, because they could provide no cost-benefit estimate relative to the baseline.

In figure 1b we have plotted data from those few studies that included measures of concurrent task performance; that is, we have plotted response to a concurrent task as a function of reliability of the automated task.

We analyzed the data in figure 1a by computing a linear regression, regressing the integer values of cost-benefit on imperfect reliability level. This regression was done twice, once with and once without the smaller subset of conditions that in table 1 were labeled high workload.

The regression correlation for the full data set was $r = +0.63$ [$p < .01$], with reliability thereby accounting for 41% of the variance. The regression equation of automation effectiveness (E) as a function of reliability [r] was $E = .05r - 3.57$. Importantly, the predicted crossover point where $E = 0$ lies at $r = 0.70$. That is, a 0.70 reliability value is an estimate of the lower level of effectiveness of imperfect diagnostic automation. Confidence intervals around the regression line, shown in figure 1, indicate that the 95% confidence estimate of this crossover is between $r = .63$ and $r = .77$.

When the regression was then re-computed with the high workload studies removed, the data fit degraded considerably. The correlation was reduced to a value of $r = +0.47$ ($p = .04$), and the regression slope itself was reduced in its value, compared to the value when the high workload points were included. This finding suggests that the dependence upon automation is more heavily manifest in high workload. To confirm this point, we re-ran the regression analysis only on the eleven high workload data points. The correlation is quite high ($r = +0.76$), and in spite of the much smaller sample size, the relation remains highly significant ($p < .01$).

Finally, the data points in figure 1b, representing concurrent task performance, were subjected to a similar regression analysis, but failed to yield any trend ($r = +0.32$, $p > .10$), suggesting that, across the current studies, performance on concurrent tasks is little influenced by the reliability of automated tasks, a finding to be discussed below.

DISCUSSION

The objective of this study was to explore the implications of varying levels of unreliability or imperfection of diagnostic automation. Our analysis suggests first, that performance is quite sensitive to the level of imperfection, suggesting, not surprisingly, that humans do not fully compensate for the greater imperfections by allocating more attention to the raw data. Rather, the joint human-automation performance appears to be pulled down by the increasing imperfections of the automation, to the extent that, below a level of around 0.70, diagnostic monitoring is worse than had the human not used the automation at all (e.g., the baseline condition). It is interesting to note that the only other report to apparently have offered up such a reliability level “constant” (in the form of a range of values), is the review of the literature carried out by Lee and See (2004), who offer that “...some evidence suggests that below a certain level of reliability, trust declines quite rapidly. The absolute level of this drop off seems to be highly system and context dependent, with estimates ranging from 90% and 70% to 60%” [p. 72]. That is, their qualitative observation identifies a range of values overlapping those observed here.

In considering the current results, we can inquire as to the source of this somewhat disconcerting downward pull of bad automation, akin to holding onto a cement life preserver in the water. Why can't/don't operators simply ignore it, and rely upon their own perceptual/diagnostic processes, as they do in the baseline condition? It appears that the answer is not that operators are unaware of the imperfections, the explanation behind the “complacency effect” typically shown in the response to first failures (Parasuraman et al. 1993, Yeh et al. 2003). We reject this explanation in the current context because the studies chosen adhered to criterion 4 above where operators were signaled as to the existence of unreliability (if not always the precise value of that unreliability).

Instead, our analysis suggests that operators chose to depend on the imperfect automation, knowing that it is far from perfect, in order to preserve available processing resources for other tasks. Two aspects of the data support this conclusion. First, the dependency relationship is stronger in high workload studies (where resources are more at a premium) than those where workload was not explicitly made high. Secondly, as shown in figure 1b, performance of concurrent tasks, in those studies that have reported these, appears to be

relatively well “buffered” from the effects of varying imperfection level, as if operators are giving those tasks whatever resources are required to preserve a constant performance level

($r = 0$), and instead, allow the degrading reliability to increasingly disrupt performance of the automation task. (The latter conclusion must be offered tentatively because the relatively low statistical power – $N = 11$ -- does not allow great confidence in confirming the null hypothesis of no effect).

Stated in other terms, operators do not appear to be aggressively re-allocating more perceptual resources to processing the “raw data” of the diagnostic task as automation’s processing of those data degrades (although Dixon and Wickens 2004a, and Wickens Dixon Goh and Hammer 2005 have found some evidence for this compensation to occur). The reason for this compensation failure (reflected in the linear slope of the function in figure 1a) even in the “non-high workload” studies, may be that the human operator has an inherent need to protect some reserve capacity for unexpected tasks that may arise in the concurrent task domain.

The compensation failure may also be because, across the dual tasks studies, humans have inherently treated automated tasks as “secondary” and therefore allocated necessary resources to the “primary” non-automated task, in order to sustain high performance. Recent data in an experiment when task priorities were explicitly manipulated supports this view: when the automated task was designated primary, the concurrent task was much more drastically degraded by imperfect automation (Wickens and Dixon, 2005).

As human factors practitioners consider using the current results to formulate guidelines, it is probably too simplistic to advise designers that “diagnostic automation with reliability greater than 0.71 is OK, and less than 0.71 should not be used”. In addition to the fact that the constant is only an estimate (which was surrounded by a 95% confidence interval of roughly 0.14), three key aspects of the costs of imperfect diagnostic automation must be taken into account. First, as implicitly suggested above, as concurrent tasks have greater importance (costs of performance failures), compared to the automated task itself, tolerance for greater imperfection (lower reliability) can be increased. Second, we have not discriminated in the current analysis between diagnostic imperfections that create misses from those that create false alarms (although some of the studies in our analysis have explicitly done so; Dixon and Wickens 2004a, Maltz and Shinar 2003). Thus, for example, in particular applications, automation misses may have a much more serious consequence, and reliability thus degraded would require a higher reliability standard, than automation degraded by false alerts. Third, we acknowledge the potential impact of many contextual factors related to training or to understanding the source of imperfection, that could modify the effects. These remain to be represented in a quantitative form.

Given the methodology of the current analysis, and the approximations used in coding the data of a particular study, we recognize that some inaccuracies may have occurred. We have, for example, glossed over effect size, by our very coarse 3-level categorization, and we did not try to quantify workload. Also, on occasion, we may have misrepresented the true reliability of a given study, where we have tried to estimate this value from other parameters reported in the experimental write up. Fortunately however, the number of studies is great enough that we believe the general conclusion is relatively robust to a few such errors.

Finally, we refer back to one critical element that restricted our coverage: the focus on imperfect diagnostic automation (stages 1 and 2) rather than, for example, stages 3 and 4 decision aids (incorporating values and costs of decision outcomes). While we acknowledge that the overall contributions of this report could be of increased value were these to be modeled (and we hope that other researchers will do so; see Parasuraman and Sheridan, 2001 for one approach), we also suspect that the methods used to represent reliability in these cases will be harder to quantify in a single universal metric, perhaps thwarting the ability to arrive at a meaningful “constant”. Just as the range of human abilities is so large as to hinder any single “theory” or “model” of human performance limits, so we suspect there to be corresponding limitations for a single all encompassing quantitative model of human dependence on imperfect automation.

ACKNOWLEDGMENTS

This research was supported by a grant # ARMY MAD 6021.000-01 from Micro Analysis and Design. David Dahn was the scientific technical monitor.

REFERENCES

- BLISS, J. and ACTON, S. A. 2003, Alarm mistrust in automobiles: How collision alarm reliability affects driving, *Applied Ergonomics*.
- DAVISON, H. J. and WICKENS, C. D. 2001, Rotorcraft hazard cueing: The effects on attention and trust, *Proceedings of the 11th International Symposium on Aviation Psychology* (Columbus, OH: The Ohio State University).
- DINGUS, T. A., MCGEHEE, D. V., MANAKKAL, N., JAHNS, S. K., CARNEY, C. and HANKEY, J. M. 1997, Human factors field evaluation of automotive headway maintenance/collision warning devices, *Human Factors*, 39, 216-229.
- DIXON, S., MCCARLEY, J., and WICKENS, C.D. 2005, Miss-prone vs. false-alarm-prone automation, Technical Report AHFD-05-16/MAAD 05-4, University of Illinois, Aviation Human Factors Division, Savoy, IL.
- DIXON, S. and WICKENS, C.D. 2004a, Automation reliability in unmanned aerial vehicle flight control, *Proceedings of the 5th Human Performance, Situation Awareness and Automation Technology Annual Meeting*.
- DIXON, S. R. and WICKENS, C. D. 2004b, Reliability in automated aids for unmanned aerial vehicle flight control: Evaluating a model of automation dependence in high workload, Technical Report AHFD-04-5/MAAD-04-1, University of Illinois, Aviation Human Factors Division, Savoy, IL.
- DIXON, S., WICKENS, C. D. and CHANG, D. 2004, Unmanned aerial vehicle flight control: False alarms versus misses, *Proceedings of the 48th Annual Meeting of the Human Factors and Ergonomics Society* (Santa Monica, CA: Human Factors and Ergonomics Society).
- DZINDOLET, M. T., PIERCE, L. G., POMRANKY, R., PETERSON, S. and BECK, H. 2001, Automation reliance on a combat identification system, *Proceedings of the 45th Annual*

- Meeting of the Human Factors and Ergonomics Society* (Santa Monica, CA: Human Factors and Ergonomics Society), 532-536.
- GALSTER, S. M., BOLIA, R. S., ROE, M. M. and PARASURAMAN, R. 2001, Effects of automated cueing on decision implementation in a visual search task, *Proceedings of the 45th Annual Meeting of the Human Factor Society* (Santa Monica, CA: Human Factors and Ergonomics Society), 321-325.
- LEE, J. D. and SEE, K. A. 2004, Trust in automation: Designing for appropriate reliance, *Human Factors*, 46, 50-80.
- LEHTO, M. R., PAPASTAVROU, J. D., RANNEY, T. A. and SIMMONS L. A. 2000, An experimental comparison of conservative versus optimal collision avoidance warning system thresholds, *Safety Science*, 36-3, 185-209.
- MALTZ, M. and MEYER, J. 2001, Use of warnings in an attentionally demanding detection task, *Human Factors*, 43(2), 217-226.
- MALTZ, M. and SHINAR, D. 2003, New alternative methods in analyzing human behavior in cued target acquisition, *Human Factors*, 45(2), 281-295.
- MALTZ, M. and SHINAR, D. 2004, Imperfect vehicle collision warning systems can aid drivers, *Human Factors*, 46(2), 357-366.
- METZGER, U. and PARASURAMAN, R. 2005, Automation in future air traffic management: Effects of reliable and imperfect detection aids on controller performance and workload, *Human Factors*, 47.
- MEYER, J. 2001, Effects of warning validity and proximity on responses to warnings, *Human Factors*, 43(4), 563-572.
- MILLER, G. 1956, The magical number 7 plus or minus two: Some limits on our capacity for processing information, *Psychological Review*, 63, 81-97.
- MOLLOY, R. and PARASURAMAN, R. 1996, Monitoring an automated system for a single failure: vigilance and task complexity effects, *Human Factors*, 38, 211-322.
- PARASURAMAN, R., MOLLOY, R. and SINGH, I. L. 1993, Performance consequences of automation-induced "complacency", *International Journal of Aviation Psychology*, 3(1), 1-23.
- PARASURAMAN, R., SHERIDAN, T. B. and WICKENS, C. D. 2000, A model for types and levels of human interaction with automation, *IEEE Transactions on Systems, Man, and Cybernetics*, 30(3), 286-297.
- SHERIDAN, T. 2002, *Humans and automation: System design and research issues*. (Wiley Interscience).
- SKITKA, L., MOSSIER, K. and BURDICK, M. 1999, Does automation bias decision-making? *International Journal of Human-Computer Studies*, 51(5).
- SORKIN, R. D. and WOODS, D. D. 1985, Systems with human monitors, a signal detection analysis, *Human-Computer Interaction*, 1, 49-75.

- ST. JOHN, M. and MANES, D. I. 2002, Making unreliable automation useful, *Proceedings of the Human Factors and Ergonomics Society 46th Annual Meeting* (Santa Monica, CA: Human Factors and Ergonomics Society).
- SWETS, J. A. and PICKETT, R. M. 1982, *The evaluation of diagnostic systems*. (New York: Academic Press).
- THOMAS, L. C., WICKENS, C. D. and RANTANEN, E. 2003, Imperfect automation in aviation traffic alerts: A review of conflict detection algorithms and their implications for human factors research, *Proceedings of the 47th Annual Human Factors Ergonomics Society Conference* (Santa Monica, CA: Human Factors and Ergonomics Society).
- WICKENS, C. D. 2000, Imperfect and unreliable automation and its implications for attention allocation, information access and situation awareness, Final Technical Report ARL-00-10/NASA-00-2, University of Illinois, Aviation Research Laboratory, Savoy, IL.
- WICKENS, C.D., and DIXON, S. 2005, Task priorities and imperfect automation, Technical Report AHFD-05-17/MAAD-05-5, University of Illinois, Aviation Human Factors Division, Savoy, IL.
- WICKENS, C.D., DIXON, S. and AMBINDER, M. in press, Computational models of human performance in UAV operation. In H. Pederson and N. Cooke. Human Factors of Remotely Operated Vehicles,
- WICKENS, C. D., DIXON, S., GOH, J. and HAMMER, B. 2005, Pilot dependence on imperfect diagnostic automation in simulated UAV flights: An attentional visual scanning analysis, Technical Report AHFD-05-2/MAAD-05-2, University of Illinois, Aviation Human Factors Division, Savoy, IL.
- WICKENS, C. D., GEMPLER, K. and MORPHEW, M. E. 2000, Workload and reliability of predictor displays in aircraft traffic avoidance, *Transportation Human Factors Journal*, 2(2), 99-126.
- WIEGMANN, D. A., MCCARLEY, J., WICKENS, C. D. and KRAMER, A. F. 2003, Effects of age on utilization and perceived reliability of an automated decision-making aid for luggage screening, *Human Factors* (accepted pending revision).
- XU, X., WICKENS C. D. and RANTANEN, E. 2004, Imperfect conflict alerting systems for the cockpit display of traffic information, Technical Report AHFD-04-08/NASA-04-02, University of Illinois, Aviation Human Factors Division, Savoy, IL.
- YAACOV, A. B., MALTZ, M. and SHINAR, D. 2003, Effects of an in-vehicle collision avoidance warning system on short- and long-term driving performance, *Human Factors*, 44(2), 335-342.
- YEH, M., MERLO, J. L., WICKENS, C. D. and BRANDENBURG, D. L. 2003, Head up versus head down: The costs of imprecision, unreliability, and visual clutter on cue effectiveness for display signaling, *Human Factors*, 45(3), 390-407.
- YEH, M., and WICKENS, C. D. 2001, Display signaling in augmented reality: The effects of cue reliability and image realism on attention allocation and trust calibration, *Human Factors*, 43(3), 355-365.