

AFCAPS-IR-2020-0004

**Individual Differences in
Multiple-Task Performance:
Status and Issues**



Oct 2020

Diane Damos, Ph.D.

Damos Aviation Services, Inc.

Prepared for:

Katie Gunther, Ph.D.

**Air Force Personnel Center
Strategic Research and Assessment
Branch**



Air Force Personnel Center
Strategic Research and Assessment
HQ AFPC/DSYX
550 C Street West, Ste 45
Randolph AFB TX 78150-4747

Approved for Public Release. Distribution Unlimited.

UNCLASSIFIED

NOTICE

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

This report was cleared for release by HQ AFPC/DSYX Strategic Research and Assessment Branch (SRAB) and is releasable to the Defense Technical Information Center.

This report is published as received with minor grammatical corrections. The views expressed are those of the authors and not necessarily those of the United States Government, the United States Department of Defense, or the United States Air Force. In the interest of expediting publication of impartial statistical analysis of Air Force tests SRAB does not edit nor revise Contractor assessments appropriate to the private sector which do not apply within military context.

Federal Government agencies and their contractors registered with Defense Technical Information Center should direct request for copies of this report to:

Defense Technical Information Center - <http://www.dtic.mil/>

Approved for public release, unlimited distribution by AFPC/DSYX Strategic Research and Assessment Branch (SRAB) Joint Base San Antonio-Randolph AFB, TX 78150-4747 or higher DoD authority. Please contact

AFPC/DSYX Strategic Research and Assessment Branch (SRB) with any questions or concerns with the report.

This paper has been reviewed by the Air Force Center for Applied Personnel Studies (AFCAPS) and is approved for publication. AFCAPS members include: Senior Editor Dr. Thomas Carretta AFMC 711 HPW/RHCI and Dr. Imelda Aguilar HQ AFPC/DSYX.

REPORT DOCUMENTATION PAGE				<i>Form Approved</i> OMB No. 0704-0188	
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.					
1. REPORT DATE (DD-MM-YYYY) 31-10-2019		2. REPORT TYPE Interim		3. DATES COVERED (From - To)	
4. TITLE AND SUBTITLE Individual Differences in Multiple-Task Performance: Status and Issues				5a. CONTRACT NUMBER HCaTs GS02Q17DCR0008	
				5b. GRANT NUMBER	
				5c. PROGRAM ELEMENT NUMBER ITSS ID10180037	
6. AUTHOR(S) Diane Damos, Ph.D.				5d. PROJECT NUMBER	
				5e. TASK NUMBER 47QFAA18F0043	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) HQ AFPC/DSYX Strategic Research and Assessment Branch				8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) Air Force Personnel Center Strategic Research and Assessment Branch JBSA-Randolph AFB TX, 78150				10. SPONSOR/MONITOR'S ACRONYM(S)	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S) AFCAPS-IR-2020-0004	
12. DISTRIBUTION / AVAILABILITY STATEMENT Approved for Public Release. Distribution is Unlimited.					
13. SUPPLEMENTARY NOTES					
14. ABSTRACT Multiple-task tests have been known to improve the predictive validity of pilot selection batteries since at least the 1930's. Despite their longstanding and widespread use for pilot selection, it is unclear what these test measure. The first goal of this report is to review the literature on individual differences in multiple-task performance. This will aid investigators in selecting the best task combinations for a pilot selection battery. The second goal is to review topics that have caused problems in assessing multiple-task performance. These include practice, task prioritization, response strategies, and data analysis. Each topic is described in detail and methods for reducing or eliminating the associated problems are given.					
15. SUBJECT TERMS Multiple-task, timesharing, parallel processing, task switching, feedback, differential stability, task prioritization, prospective memory, response strategy, control theory, individual differences, divided attention, general timesharing ability					
16. SECURITY CLASSIFICATION OF: Unclassified			17. LIMITATION OF ABSTRACT U	18. NUMBER OF PAGES 49	19a. NAME OF RESPONSIBLE PERSON Katie Gunther, Ph.D.
a. REPORT U	b. ABSTRACT U	c. THIS PAGE U			19b. TELEPHONE NUMBER (include area code) 210-565-5245

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

TABLE OF CONTENTS

Introduction	1
General Timesharing Ability	2
Background	2
Identifying a General Timesharing Ability	3
U.S. research.....	3
Croatian research.....	6
Timesharing and Personality	7
Comments	8
Individual Differences in Processing Styles	8
Response Strategy	9
Comments	14
Prospective Memory	14
Comments	15
Practice	16
Determining Stability	16
Practice to an asymptote.....	17
Reliability	18
Differential stability.....	18
Single- Versus Multiple-Task Practice.....	20
Comments	22
Controlling Task Prioritization Through Feedback	22
Which Measures Should Be Presented?	23
Terminal Feedback.....	23
Concurrent Feedback.....	24
Single-task usage.....	25
Distraction.....	26
Comments	26
Controlling response strategies	27
Restricted Response Options.....	28
Comments	29
Analysis issues	30

Reducing the Number of Dependent Variables	30
Composite measures	30
Averaged z scores.	31
Correlated common measures	31
Eliminating the Effects of Single-Task Performance	32
Multivariate analysis of covariance	32
Partialling single task.	33
Control Theory Measures	35
Comments	35
Conclusions	36
Recommendations	38
References.....	40

Individual Differences in Multiple-Task Performance: Status and Issues

INTRODUCTION

Multiple-task tests have been known to improve the predictive validity of the US pilot selection process since at least the mid-1930s (Mashburn, 1934). They were used in World War II for aircrew selection (Melton, 1947) and continued in use until 1955 (Carretta, 1990). They were reintroduced into the USAF pilot selection battery in 1993 (Carretta, Zelenski, & Ree, 2000) and have remained in the battery, where they continue to provide additional predictive validity beyond paper-and-pencil cognitive tests (Carretta, 2011, 2013). Additionally, performance on multiple-task tests has been shown to have predictive validity for civilian ab initio pilot selection (Damos & Lintern, 1981; LaRoche, 2011) as well as for both ab initio and experienced airline pilot selection (Bartram & Baxter, 1996; Doat, 1995; Huelmann & Oubaid, 2004; Trankell, 1959). They also have demonstrated predictive validities for pilot selection of foreign militaries (Hoermann, Noser, & Stelling, 2018).

Despite their longstanding and widespread use for pilot selection, it is unclear what these tests measure. Historically, the task combinations have been assumed to measure a general timesharing ability. If such an ability exists, then candidates need only to be tested on one task combination. However, as will be seen, there is very little support for a general timesharing ability.

This report has two goals. The first goal is to review the literature on individual differences in multiple-task performance. Historically, psychologists have taken two approaches to understanding individual differences in multiple-task performance. The oldest approach attempts to demonstrate the existence of a general timesharing ability. The more recent approach recognizes individual differences in multiple-task performance but attributes them to differences in fine-grained aspects of performance, such as the amount of parallel processing or task switching speed. These fine-grained behaviors have yet to be used for selection, but the underlying assumption is that they can help identify those individuals who will perform certain jobs or task combinations well. The literature on prospective memory also was reviewed as a potential new source of reliable individual differences.

The second goal is to review topics that have caused the most problems in assessing multiple-task performance. These include practice, controlling task prioritization, and data

analysis. Here, the aim is to inform investigators of common pitfalls in multiple-task performance research so such pitfalls can be avoided in future research on pilot selection (see Damos, 1991 for a discussion of common problems).

Each section of this report concludes with comments that summarizes the results and point out difficulties in interpretation. Important questions that were not answered by the research also are discussed. The final section includes a general overview of the state of multiple-task performance as it pertains to selection and a list of recommendations to improve USAF pilot selection.

GENERAL TIMESHARING ABILITY

Background

Some discussion about the nature of timesharing abilities is appropriate at the start of this section. The studies in this section attempted to identify a *general timesharing ability*. This term is rarely defined. For the purposes of this report, a general timesharing ability is an ability that affects the performance of any task combination but does not affect the performance of the component tasks. That is, it is involved exclusively in situations where the operator must perform at least two tasks concurrently.

Visual scanning is sometimes considered to be a timesharing ability. Investigators have studied experienced pilots' scan patterns to aid in the design and placement of aircraft instruments and to train less experienced pilots (Harris & Spady, 1985; Spady, Jones, Coates, & Kirby, 1982). However, complex scanning behavior such as that performed by a pilot is rarely examined in laboratory studies of timesharing abilities although display separation might be included as an independent variable in some studies (e.g. Braune & Wickens, 1986). Visual separation of the task displays has a deleterious effect on multiple-task performance because it forces the participant to move their eyes from one display to the other, adding a delay to the encoding of visual information. Display separation also might force the participant to adopt a more sequential processing strategy rather than one that involves more parallel processing (see Braune & Wickens, 1986). Depending on the number of displays to be scanned and the statistical relation between the stimuli, display separation also might adversely affect multiple-task performance by requiring the participant to develop and use an additional skill--complex visual scanning—to perform the task combination. For the purposes of this report, complex

visual scanning such as that performed by experienced pilots in flight will be treated as a distinct skill, not as a timesharing ability, and will not be discussed further.

Identifying a General Timesharing Ability

U.S. research. Beginning with Sharp (1899) many psychologists have sought to demonstrate the existence of a general timesharing ability. Work conducted prior to 1918 was hampered by a lack of statistical methods and by a limited understanding of the appropriate experimental methodology. Consequently, interpreting the results of these studies is problematic at best. The most sophisticated study of this early period (McQueen, 1917) suffered from methodological issues and limited statistical techniques. Nevertheless, the fact that McQueen failed to find any evidence of a general timesharing ability is worth noting.

After McQueen, few investigators were concerned with identifying a general timesharing ability until the late 1970s. Ackerman, Schneider, and Wickens (1984) reviewed the methodological and statistical problems commonly found in studies of a general timesharing ability. They concluded that three of four studies (Hawkins, Rodriguez, & Reicher, 1979; Jennings & Chiles, 1977; Sverko, 1977) had shortcomings that were too substantive to allow any meaningful conclusions to be drawn. However, they reanalyzed the data from the fourth study (Wickens, Mountford, & Schreiner, 1981) using a Procrustes procedure rather than a varimax procedure and limiting the number of factors to be extracted. The authors interpreted the reanalysis as providing evidence of a general timesharing ability.

The Sverko (1977) and Wickens et al. (1981) data were reanalyzed by Bittner and Damos (1986). Bittner and Damos first partialled all of the single-task measures out of all of the dual-task measures (see the Analysis Issues section for more details). They argued that this step was appropriate because the single-task measures are often significantly correlated with each other. Thus, this step removed all of the single-task contributions to dual-task performance, allowing the variance attributed to timesharing factors to be identified more clearly. Bittner and Damos then performed a principal component analysis with an eigenvalue cutoff and varimax rotation on the partialled correlations. Neither factor analysis showed evidence of a general timesharing ability. Nevertheless, the first unrotated factor in each analysis accounted for a large percentage of the variance. In a later paper, Bittner and Damos (1988) suggest that this first factor may reflect individual differences in a task switching ability. Such an ability would not preclude the existence of more specific timesharing abilities.

Brookings and Damos (1991) reviewed eight studies that used either a factor analytic approach or an interpretation of intra- and intertask correlations to identify a general timesharing ability. Four of the eight studies they reviewed had been previously discussed by Ackerman et al. (1984). Like Ackerman et al., Brookings and Damos found that these four studies, as well as two others (Fogarty, 1987; Sverko, Jerneic, & Kulenovic, 1983), had too many methodological and statistical issues to allow an unequivocal interpretation of the results.

The remaining two studies (Braune & Wickens, 1986; Brookings, 1990) merit more detailed examination. Braune and Wickens used a componential approach to identify a general timesharing ability. They identified five information processes that could contribute to multiple-task performance: parallel processing, task switching, display separation, processing flexibility, and an internal model of the statistical relation between the task stimuli. Under dual-task conditions the participants performed two, 1-dimensional compensatory tracking tasks. The distance between the tasks, control order of the tasks, and statistical relation between the inputs of the tasks were varied to assess the proposed processes. One of the strengths of this study is its treatment of display separation. Here it was manipulated as an independent variable to allow its effect on dual-task performance to be examined. The authors hypothesized that separating the displays might force participants to adopt a serial processing strategy.

The authors calculated difference scores between the various experimental conditions as measures of the five processes. Predictably, scanning had a deleterious effect on performance. Factor analyses performed on the difference scores showed a bipolar factor with significant loadings on the parallel processing variables and the scanning variables. Braune and Wickens (1996) interpreted this factor as indicating that participants who could parallel process well were poor serial processors and vice versa. A second factor was defined by variables reflecting four of the five information processes that the authors identified previously as affecting multiple-task performance. This factor was interpreted by the authors as a general timesharing factor.

Braune and Wickens (1996) drew two conclusions from their results. The first was that individual differences in dual-task performance exist, and they are related to parallel versus serial processing. The second concerns evidence for a general timesharing ability. Braune and Wickens believed that they had found evidence for such an ability (the second factor) but within a limited context (dual-task tracking). They suggested that the timesharing ability domain had a hierarchical structure with many specific timesharing abilities at the lowest level, fewer abilities at the intermediate level, and one general ability at the highest level.

Brookings (1990) developed an experiment based on Wickens's Multiple Resources Model (Wickens, 1980, 1992). He tested all possible combinations of four computerized tasks (Sternberg memory search, grammatical reasoning, probability monitoring, and unstable tracking). The probability monitoring task warrants some description. This task required the participant to monitor an electromechanical dial with a pointer that moved randomly between the two halves of the dial face every 200 ms. Occasionally, the pointer would remain in one half of the dial face 95% of the time. The participant pressed a key when he determined that the movement of the pointer was confined to half of the dial face. Brookings also administered a psychometric ability battery consisting of eight paper-and-pencil tests.

Participants performed the computerized tasks over a 2-day period. They received a total of 12 minutes of practice on each task alone and 12 minutes on each combination. The 2-day training period was divided into three sessions, with the data for analysis collected in the second and third sessions. Brookings gave no information on what, if any, feedback was provided to the participants either during or after a trial.

Brookings (1990) found that all of the dual-task data were skewed. Consequently, all of the data were transformed to normalized standard scores. Three sets of analyses were conducted using the normalized scores, the sum of the scores from the two tasks, and the product of the scores. Two confirmatory factor analyses were performed on each set of data. The first fit a model with four correlated single-task factors. The second used the four correlated single-task factors and an orthogonal timesharing factor. The five-factor analysis performed on the sum of the normalized scores did not converge. The analyses performed on the other two data sets showed a better fit with a five-factor solution than a four-factor solution. However, in both data sets some of the dual-task scores did not load significantly on the timesharing factor or loaded negatively.

Brookings (1990) concluded that this study produced no evidence for a general timesharing ability but did identify a second-order timesharing factor defined by the probability monitoring task. However, the data showed several problems that make any conclusions about a timesharing factor problematic. First, the single-task response time for the probability monitoring task increased with practice rather than decreased. Second, performance on two of the three tasks paired with probability monitoring improved with practice, whereas performance (response time) on the probability monitoring task itself did not. This result could indicate an asymmetric trade-off strategy favoring the other tasks. Brookings suggested this explanation, but the likelihood that this trade-off occurred cannot be evaluated because Brookings provided

no information about feedback. The situation is clouded further by the fact that the dual-task false alarm rate on the probability monitoring task fell from the second to the third session, which might indicate that the participants changed their within-task performance strategy over time. Finally, combinations including the probability monitoring task might have required more scanning than the other combinations. Brookings provided no information about the size of the dial for this task nor did he give any details about the separation of the tasks under dual-task conditions. Consequently, this “timesharing” factor may be related to scanning.

Croatian research. Much research on timesharing has been conducted by Sverko and his associates at the University of Zagreb. Two of the studies (Sverko, 1977; Sverko et al., 1983) were discussed only briefly earlier because of statistical issues. However, the Sverko et al. (1983) study, as well as two more recent studies (Jerneic, 2007; Sverko, Maslic-Sersic, Jerneic, & Gurdulic-Sverko, 1994), used a unique methodological approach to identify a timesharing ability that bears examination. Because the methodologies in these two studies are almost identical, only the most recent one by Jerneic (2007) will be discussed in some detail.

Jerneic (2007) addressed many of the practice, feedback, and statistical concerns of Ackerman et al. (1984) and Brookings and Damos (1991). He used three 4-alternative choice reaction time tasks that differed in their stimuli (letters, lights, and numbers), two versions of a 1-dimensional pursuit tracking task, and two versions of a 1-dimensional compensatory tracking task. These seven tasks were used to construct 14 dual-task combinations (no choice reaction time/tracking combinations). The tasks were displayed so that little, if any, scanning was required to perform the task combinations. Jerneic hypothesized that a general timesharing skill might be evident only after substantial practice. Consequently, the 111 participants completed one testing session per day for 10 days. By the end of the study, each participant had performed each individual task for 20 minutes and each combination for 30 minutes

Extensive statistical analyses were conducted on the data, including principal component analysis, principal axis factor analysis, and maximum likelihood analysis. None of the statistical analyses showed evidence of a general timesharing ability although the same three specific timesharing factors were identified in each analysis. One timesharing factor was defined by combinations composed of the choice reaction time tasks, one by combinations including the pursuit tracking task, and one by combinations including the compensatory tracking task. However, the underlying single tasks generally loaded on the corresponding timesharing factor, that is, single-task pursuit tracking loaded on the pursuit tracking timesharing factor. Consequently, Jerneic (2007) conducted another set of factor analyses on the dual-task

correlations after single-task performance had been partialled out. Jerneic again found three factors with one factor corresponding to choice reaction time task combinations and two factors corresponding to the two types of tracking tasks. He concluded, like Braune and Wickens (1986), that these were three specific timesharing factors.

Despite the large amount of practice given to the participants and the extensive statistical analyses, several aspects of this study suggest that one should exercise caution in accepting the results. One of these aspects is practice. A comparison of the intermediate and late stages of practice shows better average performance for all of the tasks in the late stage. As pointed out by Ackerman et al. (1984), practice must stabilize to some degree before an investigator attempts to identify a timesharing factor. It is unclear if such stabilization occurred in the Jerneic (2007) study. The second major concern involves dual-task methodology. The dual-task choice reaction time tasks were forced paced with an interstimulus interval of 2 seconds. The stimuli were presented simultaneously and the dependent measure—correct reaction time—was recorded only when the participant responded correctly to both stimuli within the interstimulus interval. The goal of this scoring method was to encourage the participants to timeshare (parallel process) rather than to switch between tasks. The degree to which this requirement was successful is unknown because no fine-grained analyses were performed, and no error rates or omission rates were reported. Similarly, under dual-task conditions, the percentage of time on target was recorded only when both cursors were on the target. It is important to note that Jerneic explained the scoring process to the participants in detail and, to maintain their motivation, gave substantial monetary rewards for sessions in which their performance improved compared with the previous session.

Timesharing and Personality

Jerneic and Sverko (1994) examined the relation between timesharing ability, general mental ability, and personality using essentially the same tasks and methodology reported in Jerneic (2007). Personality was assessed using the Eysenck Personality Questionnaire (Eysenck & Eysenck, 1975). The authors found the same three timesharing factors as in Jerneic (2007). Of the personality measures, only the correlation between neuroticism and the timesharing factor defined by the choice reaction time tasks was significant. This correlation was positive. Of the cognitive ability measures, only perceptual speed was significantly correlated with all three timesharing factors. It was positively correlated with the timesharing factor defined by pursuit tracking and with the factor defined by discrete task combinations. It was negatively

correlated with the third timesharing factor, which was defined by compensatory tracking. However, this third factor was defined by only one task combination.

Comments

Two conclusions can be drawn from this review. The first, and perhaps most obvious, is that meeting all the methodological and statistical requirements to identify a general timesharing ability is difficult. Practice is a major obstacle in identifying a timesharing ability because only performance data that have reached some form of stabilization should be analyzed. Determining when multiple-task performance has stabilized is challenging (see the Practice section for approaches to this problem). Analyzing the large number of dependent measures obtained from task combinations presents a formidable challenge to understanding trade-off strategies and accurately representing performance (see the Analysis Issues section). To make data analysis and interpretation more tractable, some investigators have chosen to ignore errors and omissions. However, analyses based only on correct responses may be misleading.

The second conclusion is that, after almost 120 years, the evidence for a general timesharing ability is scant at best. If a general timesharing ability exists that has a substantial effect on multiple-task performance, investigators should have found more evidence of it despite the statistical and methodological challenges involved with its identification. The concept of a general timesharing ability was developed over 100 years ago and now appears to be too simplistic. Both Braune and Wickens (1986) and Brookings (1990) suggested a hierarchical structure of timesharing abilities, perhaps similar to the three-stratum structure of cognitive abilities proposed by Carroll (1993). This concept, however, has not been pursued. A different approach has been taken by other investigators who study individual differences in multiple-task performance in terms of, for lack of a better term, processing styles. This approach will be discussed in the following section.

INDIVIDUAL DIFFERENCES IN PROCESSING STYLES

The second, and more recent, approach to studying individual differences in multiple-task performance is concerned with processing strategies. Two groups of studies have used response strategies to identify those individuals who appear to be processing information from concurrent, discrete tasks in parallel versus in serial. Typically, study participants use one of three major types of response strategy. Those who use the first type do not actually timeshare. Instead, they emit a long series of responses to one task before switching to the second. Participants who use the second type switch frequently between the two tasks, whereas those

using the third type respond concurrently to both tasks. The switching strategy might be broken down further into those who alternate between the two tasks and those who emit a few responses to one task before switching to the other.

The first group of studies was conducted in the early and mid-1980s by Damos and her colleagues (Damos, 1984; Damos & Smist, 1982; Damos, Smist, & Bittner, 1983; Damos & Wickens, 1980). The second group was done approximately 30 years later by Manzey and his colleagues (Brüning & Manzey, 2018; Reissland & Manzey, 2016). These studies will be reviewed in chronological order.

Response Strategy

In their study of the training and transfer of timesharing skills, Damos and Wickens (1980) used two task combinations. The training combination consisted of a letter classification task and a one-back memory task; the transfer combination, of two identical, 1-dimensional compensatory tracking tasks. Moving bars as described in the Feedback section provided concurrent feedback during the transfer sessions. During the training session, some of the participants were seen to emit a long series of responses to one of the component tasks before switching to the other. During breaks between the training blocks, the investigator asked the participant to limit this type of behavior. Frequently, these requests had no effect on behavior.

As part of the analyses the participants' response strategy was analyzed and classified as either "simultaneous," "alternating," or "massed." A participant who responded consistently to both tasks within 99 ms. was classified as using the simultaneous response strategy. A participant who consistently made one response to a component task before responding to the other was classified as using an alternating strategy. If the participant consistently made more than two responses to one task before responding to the other, the strategy was classified as massed. Subsequent analyses showed that the simultaneous response group had the shortest dual-task correct response interval (CRI, see the Analysis Issues section) on the discrete task combination and the smallest difference between the single- and dual-task CRI. It is important to note that the three strategy groups did not differ on their single-task performance.

Because the purpose of the study was to examine transfer of timesharing skills, the dual-task tracking data were not examined extensively for differences between the strategy groups. The authors, however, did comment on two findings. All 6 of the participants who adopted a simultaneous response strategy reached the transfer criterion on the tracking combination, whereas 3 of the 10 participants using other strategies did not. The simultaneous response

group was also found to have marginally better dual-task tracking performance than participants using the other strategies.

Damos et al. (1983) examined response strategies to determine if they reflected a consistent individual difference in multiple-task performance. The same two discrete tasks—letter classification and one-back memory—used in Damos and Wickens (1980) were used in this experiment. The authors proposed that the simultaneous response strategy was the most effective strategy because it allowed for the possibility of parallel processing of some of the information. The alternating strategy was considered to be a good one because the participant could learn to switch rapidly between the tasks. The massed strategy was not even considered to be a timesharing strategy because the participant did not attempt to interweave the tasks in any manner.

On Day 1, 37 participants first completed 12, 1-minute single-task trials on each task, followed by 5 1-minute dual-task trials. After the participant left, the investigators classified the response strategy as simultaneous, alternating, or massed. Simultaneous responses to both tasks had to be emitted within 33 ms. of each other. The greatest number of sequential responses was recorded and the percentage of between-task switches was calculated by dividing the number of actual switches by the number possible. Identification of the groups was straightforward with participants who used the simultaneous response strategy making no fewer than 71% of their responses simultaneously. None of those using the alternating strategy strictly alternated less than 85% of the time.

For Day 2, all eight of the simultaneous response participants were asked to continue using that strategy. Of the alternating response participants, 3 were told to continue with that strategy, and the other 4 were instructed to use the simultaneous strategy (ALT-SIM). Of the massed response strategy participants, 12 were asked to change to the alternating strategy (MASS-ALT), and 10 to the simultaneous strategy (MASS-SIM). After receiving two 1-minute single-task trials, the participants performed five dual-task blocks of five trials. Each block was followed by one single-task trial on each task. Dual-task data were averaged over the two tasks and analyzed by response strategy. The Group X Trial ANOVA performed on the three groups using the simultaneous response strategy showed a main effect of group and a Group X Trial interaction, with the MASS-SIM group performing more poorly than the other two. Curve-fitting revealed that the MASS-SIM group's estimated asymptote was significantly higher than those of the other two groups. ANOVAs on the single-task data revealed no significant between-group differences. Rough estimates of parallel processing were obtained on Day 2 by averaging the

single-task CRIs for both tasks on the trials that preceded and followed a dual-task block and subtracting the average from the average dual-task trial on that block. The MASS-SIM group showed less evidence of parallel processing than the other two groups although the between-group difference was not statistically significant. Comparable analyses performed on the alternating response groups showed no main effect of group on averaged dual-task performance. However, estimates of switching time showed that the ALT-ALT group switched between tasks significantly faster than the MASS-ALT group.

These results suggest that individuals who adopted the massed response strategy perform more poorly under dual-task conditions than those who selected either of the other two strategies. This study and that of Damos and Wickens (1980) provide no information about the generality of these results or their reliability. They also provide no information about any demographic factors that might account for the choice of response style.

Damos and Smist (1982) did a small study examining the relation between response strategy and performance on other dual-task combinations. Eleven participants performed the letter classification task and the one-back memory task used in the prior studies, a dichotic listening task, and a dual-task tracking task consisting of two, identical 1-dimensional compensatory tracking tasks. Five participants used the simultaneous response strategy, three the alternating strategy, and three the massed strategy. Only differentially stable data were analyzed. Analyses revealed no between-group differences on any single-task measure. The multivariate ANOVA (MANOVA) showed a main effect of group. The only significant univariate analysis was for the dichotic listening task, which appears to show worse performance by the alternating group. The test for the memory-letter classification task was marginally significant ($p = 0.06$), with better performance by the simultaneous response group.

Results from this study must be interpreted with caution because of the small sample size. However, a second study (Damos, 1984) had 30 participants perform a variety of information processing tasks, including complex auditory monitoring and a choice reaction time task. They also completed hidden figures, time estimation, and digit span tests. The participants performed the letter classification and one-back memory tasks using the procedure described in Damos et al. (1983). Of the 30 participants 5 did not show a clear response pattern and were classified as “mixed.” The participants gave subjective workload ratings on all of the tasks and tests using a 10-dimensional rating scale.

Two of the workload scales showed significant differences, with the mixed response group reporting more difficulty and more workload than the alternating group. Digit span and time estimation were significantly correlated. Consequently, a MANOVA was performed on the data from these two tests. The main effect of group was not statistically significant ($p = 0.09$). However, the data did show that the massed response group had the lowest digit span (6.5) versus an average of 7.3 for the other three groups. The simultaneous group gave the most accurate estimate of a 10-s interval (10.5 s), whereas the massed group gave the worst (7.7 s). Any results again must be interpreted with caution because of the small sample size (10 in the simultaneous group, 8 in the alternating, 6 in the massed, and 5 in the mixed).

Manzey and his colleagues hypothesized that individuals have different approaches to multi-tasking. Some prefer a serial (successive) approach to multi-tasking, whereas others prefer to overlap some of the processing. In the first of two studies, Reissland and Manzey (2016) used the idea of a preview of task stimuli to identify the two types of multi-task processors. The authors employed a classical task switching paradigm and a variation in this study. In the classical paradigm the stimulus for only one task is displayed at a given time. After the participant responds, the stimulus for the other task is displayed. This paradigm forces the participants to process the tasks serially. In the variation of the classical paradigm Reissland and Manzey allowed a preview of the stimuli for the other task. That is, the participant had to respond to one task but could see the stimuli for the next task. The authors hypothesized that some individuals would attempt to ignore the stimuli for the next task; others, would attempt to use the information to begin processing on the next task.

Sixteen participants performed the classical task switching paradigm (no preview). Another 16 performed the variation (with preview). Three stimuli were presented for each task before requiring a switch to the other task. This pattern allowed the investigators to examine performance immediately before and after a switch, and they could examine reaction times for repeated responses to the same task. At the start of the study, the participants received practice on each task alone. The 25% quartile (fastest 25%) of a participant's responses was used as a baseline for dual-task performance. Those dual-task responses that fell into the 25% quartile were classified as "fast responses." Each participant's dual-task data were examined to determine how many fast responses the participant emitted. Fast responses can occur randomly under dual-task conditions. Consequently, the highest average percentage of fast responses (8.5%) emitted by any participant in the classical paradigm was used as a cutoff for those in the preview group. Only participants who had a percentage of fast responses higher

than the cutoff were classified as “overlapping processors.” The other participants were classified as “serial processors.” Of the preview group 9 were classified as overlapping processors (average of 27.6% fast responses) and 7 as serial processors (average of 5.6% fast responses). The estimated average switching time for participants in the classical paradigm did not differ from that of the serial processors and was approximately 250 ms. The comparable estimated switching time for the overlapping processors was 50 ms.

In the second of two studies Reissland and Manzey (2016) used the same two tasks. They told the study participants that the tasks were equally important and to respond as best as they could. Thus, the participants had to develop their own strategy to maximize their performance. The participants always had a preview available of the other task. Participants who switched less than 10% of the time were called “blockers” and correspond closely to the massed strategy participants of Damos and her colleagues. Of the 17 participants, 6 were classified as blockers and switched, on average, less than 2% of the time. The second group ($N = 6$), “switchers”, switched between the tasks about 21% of the time, whereas the “alternators” switched approximately 93% of the time. Using fine-grained analyses and the concept of the fast response as defined in Experiment 1, the authors were able to determine that the switchers performed some parallel processing using the preview. The alternators, on the other hand, showed no evidence of parallel processing.

The authors proposed that blockers actively avoid multiple-task situations and attempt to protect performance on the current task as much as possible from the processing demands of the second task. The alternators also might prefer serial processing. However, they adopt an alternating strategy because it avoids real-time decision-making about when to switch tasks and subjectively might reflect compliance with the equal priority instructions. The authors also remarked that the “cautious” strategy included in some computational models might reflect the preference for serial processing, whereas the “risky” strategy might refer to parallel processing.

A follow-up study by Brüning and Manzey (2018) used essentially the same paradigm as Reissland and Manzey (2016) but focused more on inter-task switching speeds. Additionally, participant’s working memory was assessed prior to performing the experimental tasks. The authors identified a small group of participants ($N = 4$ of 45) who demonstrated overlapping processing under all experimental conditions. A second group used overlapping processing under some conditions and serial processing under others. The serial processors used serial processing under all experimental conditions ($N = 18$ of 45). This demonstrates that processing strategies are flexible in some people and complements the Damos et al. (1983) study.

Working memory appeared to be related to processing strategy. That is, participants with a greater span were more flexible in their processing strategy than those with a lower span.

Comments

Most research on multiple-task performance has ignored individual differences, considering them to be a nuisance rather than a topic for investigation. As a result, only two groups of studies were found that specifically investigated individual differences in multiple-task performance. Both used discrete task combinations and examined participants' response strategies. Damos and her colleagues (1983) attempted to determine if the different response strategies reflected innate cognitive processing styles that are not trainable. Manzey and his colleagues addressed the same questions but focused on fine-grained analyses of switching behavior. Both sets of studies support the idea that the response strategies reflect consistent individual differences in cognitive processing, but that individuals have some flexibility in how they process information. The most recent study (Brüning & Manzey, 2018) suggested that the response strategy might be related to working memory capacity, a potentially useful result for selection and classification. To date, however, no research has attempted to replicate this result.

PROSPECTIVE MEMORY

As noted by Dismukes (2012) the term "prospective memory" is somewhat misleading. It does not refer to a distinct memory system, such as working memory. Instead, it refers to an intention to perform a task at a later time. A defining characteristic of prospective memory is the lack of a cue triggering the intended action. That is, the operator must remember to remember. Prospective memory does not appear to be a distinct attribute; instead, it involves a variety of attributes, including task management, planning, and attention (Dismukes, 2012, p. 215).

Prospective memory is important because interruptions occur frequently in flight. A classic example is a pilot is performing a checklist when he or she is interrupted by another, higher priority task. The pilot must form a prospective memory of the incomplete checklist to return to it and complete it after the interrupting task has been completed. Failures of prospective memory are known to have serious consequences when they occur in workplace situations such as flying an aircraft (Dismukes, 2012) and practicing medicine (Grundgeiger, Sanderson, & Dismukes, 2014).

Very few prospective memory studies have assessed the relation between cognitive abilities and performance on a prospective memory task in healthy, working-age adults. One such study (Zuber, Kliegel, & Ihle, 2016) examined the contribution of episodic memory to performance on a two-back memory task using letters as stimuli. The participants made a “yes” response if the current letter had been presented two stimuli previously. Otherwise, they responded “no.” The prospective memory task required the participant to make a different response when the current stimulus was either an A or a D. Episodic memory was assessed using the Wechsler Adult Intelligence Scale – Revised (WAIS-R) and with a delayed story recall test. Episodic memory was found to be significantly correlated with performance on the prospective memory task.

A second study (Van Benthem, Herdman, Tolton, & LeFevre, 2015) tested pilots with ratings from student to Air Transport on a series of simulated touch-and-go landings at uncontrolled airports. Participants were to make radio calls at six locations in the traffic pattern. Prospective memory performance was the percentage of radio calls made at each of the six locations in the traffic pattern. Two cognitive abilities, monitoring and working memory, were assessed using the DriveAble Cognitive Assessment Tool (DCAT), a battery designed to identify older drivers who might be at risk for operating a motor vehicle. DCAT produced age-normed z scores on the two subtasks, which were averaged. Pilots were classified as having high or low cognitive health based on a median split of the averaged z scores. Pilots with high cognitive health completed approximately 8% more radio calls (81% versus 89%) than those with low scores. The authors attributed this difference to differences in working memory. However, both DCAT tasks appear to assess several skills and abilities. Attributing the performance differences on the prospective memory task to working memory requires further confirmation.

Comments

Prospective memory has been implicated in several aviation accidents and incidents (Dismukes, Kochan, & Goldsmith, 2018; Nowinski, Holbrook, & Dismukes, 2003). Consequently, its importance in successful aircraft operations is no longer disputable. Currently, however, there appears to be little in the prospective memory literature that could be used to improve the USAF pilot selection system. Einstein, McDaniel, Williford, Pagan, and Dismukes (2003) observed that since the 1990’s, the amount of basic research concerned with prospective memory has been increasing. Much of this work, such as the Zuber et al. (2016) study, concerns the basic cognitive processes that underlie prospective memory. As such, the

tasks employed are simple and bear little resemblance to the types of tasks that are required in flying. In studies using more flight-related tasks, such as that by Van Benthern et al. (2015), the performance measure is often the percentage of prospective memory actions performed (e.g., West, Scalaro, & Bailey, 2001). Often few prospective memory events are presented. Consequently, the dependent measure might not be a reliable reflection of the participant's performance. None of the studies reviewed for this section had test-retest reliabilities for their performance measures, and no data on ethnic or gender differences were located. Nevertheless, the role of working memory in prospective memory suggests that further assessment of USAF candidates' working memory could be beneficial.

PRACTICE

Psychologists have known for many years that performance on both simple (Bradley, 1969) and complex (Fitts, 1962) tasks continues to improve even after extensive practice. Consequently, an investigator studying multiple-task performance must contend with the fact that both single-and multiple-task performance might improve over an extended period. Because multiple-task performance should reach some type of stability before the experimental factors are introduced and data collection begins (Ackerman et al., 1984), the investigator must first determine a combination of single- and multiple-task practice that allows multiple-task performance to reach stability efficiently.

This section is concerned with how single-task practice affects multiple-task performance. It begins with a discussion of three methods of identifying performance stability. For each method, one example of a study using that method is provided. Next, the literature on single- versus dual-task training is presented. Only three studies were located that specifically addressed this topic. Part- versus whole-task training is a closely related topic that includes many studies of multiple-task performance. The interested reader might wish to consult this literature for additional information. (For a review of the part- task versus whole-task literature, see Wightman & Lintern, 1985)

Determining Stability

Psychologists have used three approaches to determine how much practice is sufficient to reach performance stability. The traditional approach has been to "practice to an asymptote." The second uses inter-trial reliability measures, and the third, the concept of differential stability. These three will be discussed in turn.

Practice to an asymptote. Many studies use “practice to an asymptote” as a criterion for determining how much practice to give on the component tasks, but how an asymptote is identified is rarely discussed. In some studies, the identification appears to be based on a subjective judgment by the investigator. In other cases, the investigator performs a statistical test on the last two or three blocks of trials. If there is no statistically significant difference in performance between the blocks, the task is assumed to have reached an asymptote.

Damos (1986) provided an example of the problems that can occur with this approach. Pretests were conducted to identify the amount of practice needed to reach asymptotic performance on each of four reaction time tasks. Each task had two dependent variables: reaction time and percentage of error. The point at which performance reached an asymptotic level was identified for each task using visual inspection.

The main study included participants who were recruited from the same participant pool as for the pretest. Two dual-task combinations were used. Because both task combinations showed similar single- and dual-task results, the data from only one combination will be discussed. A three-group design was used. Group 1 received enough practice on each component task to reach single-task asymptotic performance based on the pretest data. Group 2 received 50% of the required trials, and Group 3 received one trial on each task. After 25 dual-task trials, the participants performed one trial on each component task. The scores on the single-task trials immediately preceding and following dual-task practice were compared for Groups 1 and 2. Both groups had lower percentage error scores and shorter correct reaction times after dual-task training than before but on only one of the two tasks. Under dual-task conditions, performance improved on both tasks with practice. However, the anticipated between-group differences in dual-task performance were weak; of the four dependent measures only one showed a main effect of group, and there were no significant Group x Trial interactions.

These results suggest that the amount of single-task practice had little effect on subsequent dual-task performance. The main experimental question, “Does single-task practice to an asymptote result in more rapid improvements in dual-task performance than less single-task practice” remained unanswered because Group 1 never reached single-task asymptotic performance. Clearly, the visual determination of an asymptote for one of the two tasks was incorrect. If an investigator wants to ensure that the participants have reached asymptotic

performance on the component tasks, modern curve fitting methods or the two methods described below are better alternatives than visual inspection.

Reliability. To employ this method, the investigator first calculates the inter-trial correlation matrix. The investigator then looks for an entry in the matrix after which the intercorrelations exceed a predetermined value. Although investigators rarely specify an acceptable level of inter-trial correlation, the rule of thumb seems to be 0.70. For example, if Trial 7 and all trials after Trial 7 have an inter-correlation matrix with cell entries exceeding 0.70, the investigator would consider performance beginning with Trial 7 to be stable.

North, Harris, and Owens (1978) reported one of the few studies designed explicitly to examine test-retest reliability for a task combination and its component tasks. The study used a combination of a compensatory tracking task and a choice reaction time task, a combination used frequently in multiple-task research. Twenty-six naval aviation officer candidates completed two testing sessions per day for two days. By the end of the fourth session the participants had made 400 responses to the choice reaction time task and had completed 15 minutes of single-task tracking and 12 minutes of the dual-task combination. Under dual-task conditions, the participants received concurrent feedback using the moving feedback bars described in the Concurrent Feedback section.

The correct response interval (CRI; see the Analysis Issues section) and the percentage correct responses were the dependent measures for the choice reaction time task. Root mean squared error (RMS) was calculated for the tracking task. The test-retest reliabilities were calculated between the second testing sessions of each day. Only the CRI correlations under both single- and dual-task conditions exceeded 0.7.

This study is an example of a situation in which only one of the dependent measures reached the rule-of-thumb level of reliability. High levels of test-retest reliability might require extensive single- and multiple-task practice.

Differential stability. The concept of differential stability was developed by Jones (1970a, 1970b) to describe individual differences in learning. It was latter used with repeated measures designs to detect small changes in performance caused by experimental factors. A task is considered to be differentially stable if, after X trials of practice, a) the mean level of performance no longer improves or improves in a slow, linear manner, b) the variance is constant from trial to trial, and c) the inter-trial correlations after Trial X remain constant (Jones, Kennedy, & Bittner, 1981).

Bittner(1979) suggested three ways to identify the point at which differential stability occurs. The first is a graphical method that plots the inter-trial correlations using different trials as the base. The horizontal axis can be thought of as trials after the base trial. For example, if the practice session had 15 trials, one plot would show the correlations between Trial 1 and each successive trial (14 data points). A second plot would show the inter-trial correlations between, for example, Trial 4 and each successive trial. The correlation between Trial 4 and Trial 5 would be plotted at the same point on the horizontal axis as the correlation between Trial 1 and Trial 2. Then another trial would be selected as the base trial and the points plotted.

The slopes of the plots are then compared visually. Plots that have approximately zero slopes show differential stability. The advantage of this method is that it provides an intuitive understanding of the data. Its disadvantage is that it does not provide a statistical test locating the trial on which performance became differentially stable.

The second method is called *early versus late correlation analysis of variance (ANOVA)*. With this method, the investigator first attempts to divide the inter-trial correlation matrix into an early and a late phase of practice. The late phase practice trials are the trials that are assumed to be differentially stable. Therefore, the late phase submatrix should show relatively constant correlations across the row indicating that all the late phase trials have about the same correlation with a given early phase trial. A two-way ANOVA is then conducted on the submatrix. If the row effect is non-significant, then the investigator has chosen the correct breakpoint between early and late trials. If it is significant, the investigator chooses a later trial as the breakpoint and recalculates the ANOVA until the trial at which performance became differentially stable is identified.

One drawback to this second approach is that the errors in the covariance matrix used to calculate the correlations might be correlated. This will tend to inflate the significance level. Nevertheless, for the purposes of locating differential stability, an inflated significance level will result in a conservative test because a non-significant test is the desired outcome.

The third method uses Lawley's Test for correlation equality (Lawley, 1963). To use this test, the investigator again must use the procedure described above, dividing the inter-trial correlation matrix into an early versus a late phase of practice. Lawley's test is conducted only on the late phase inter-trial correlation matrix. If the test is non-significant, the investigator has correctly identified the trial during which performance becomes differentially stable. The major drawback of using the Lawley's Test is that it assumes multivariate normality.

Damos, Bittner, Kennedy, and Harbeson (1981) provided an example of the use of differential stability for studying multiple-task performance. Their study was conducted as part of an extensive research program at the Naval Biodynamics Laboratory that examined methodological issues in repeated-measures paradigms. All studies in this program were conducted using the same testing protocol: The participants completed 15 trials of an experimental task each work day for three weeks. The participants were enlisted naval personnel assigned to the Naval Biodynamics Laboratory. Each participant typically completed several studies using the standard protocol as part of his duties. For the Damos et al. experiment, the participants performed two identical critical tracking tasks. These participants previously had completed studies in which they performed the critical tracking task alone, a compensatory tracking task, and a dual-task combination consisting of the critical tracking task and a compensatory tracking task. Thus, by the time they began this study, they had completed 225 single-task trials of the critical tracking task and 225 dual-task trials.

Despite the large amount of prior practice, performance on the critical tracking task combination continued to improve across the 15 testing days. Performance increased essentially linearly after the sixth day of testing with the standard deviation becoming constant at about the same time. The inter-trial correlations were plotted using the graphical method described above. This method suggested that differential stability was obtained on Day 10. The Lawley's Test conducted on the submatrix of correlations from Day 10 to 15 confirmed that differential stability occurred on Day 10. However, two facts are worth noting. First, the average correlation for Days 10-15 was 0.74. Thus, differential stability can occur with relatively low levels of test-retest reliability. Second, a prior study showed that single-task critical tracking performance reached differential stability on Day 6, but performance showed no evidence of reaching an asymptote. Similarly, the dual-task data from this study reached differential stability on Day 10 but showed no evidence of reaching an asymptote by Day 15. This again indicates that practice to an asymptote is an undesirable method for determining how much single-task practice should precede dual-task performance.

Single- Versus Multiple-Task Practice

Multiple-task performance is too complex a topic for a comprehensive statement about the effect of single- versus dual-task practice on performance. Perhaps for this reason only three studies on this topic were located.

Rieck, Ogden, and Anderson (1980) examined five levels of single-task practice on dual-task performance. Each of five groups had a total of four, 3-minute training trials. The groups received 0, 1, 2, 3, or 4 single-task trials and 4, 3, 2, 1, and 0 dual-task trials, respectively. All groups then transferred to two different task combinations. The training and the first transfer combination used the same tracking task but different discrete tasks. The second transfer task used both a different tracking task and a different discrete task. During the single-task training trials the participants practiced only the tracking task.

Single-task performance on the tracking task continued to improve across all training trials. The amount of single-task practice had little effect on the performance of the first dual-task trial of the training task combination for the four groups who received dual-task training. Similarly, the amount of single-task practice had no effect on the performance of the two transfer tasks. However, the amount of prior dual-task practice did affect performance on the first transfer task, but not on the second.

The results of this experiment are difficult to interpret for two reasons. The first is that the transfer trials were only 3 minutes long. This period might have been too brief to detect any between-group differences when both of the component tasks were new. The second reason is that no single-task practice was given during training on one of the component tasks. This is an unusual methodology, and the authors do not provide a rationale for its use. The groups appear to have adopted different speed/accuracy trade-offs during the first dual-task training trial and the first transfer trial. These differing speed/accuracy trade-offs make it very difficult to understand the relation between the amount of single-task practice on one component task and subsequent dual-task performance.

Damos (1987) examined differing amounts of single-task practice on two identical 1-dimensional compensatory tracking tasks. Groups 1, 2, and 3 received 1, 3, and 6 1-minute trials on each task, respectively, followed by 25 dual-task trials. The participants returned one week later for a retention test consisting of 25 dual-task trials. No between-group differences were found on the dual-task training data. The retention data showed a significant effect of group only on one of the tasks. Post hoc analyses could not identify the source of this effect.

Detweiller and Lundy (1995) examined the effect of single- versus dual-task practice on visual search tasks with consistent mapping (see Shiffrin & Schneider, 1977 for a discussion of consistent mapping). One task was a word-category search task; the other, a spatial-pattern search task. Experiment 1 consisted of two training sessions and one transfer session. Both

tasks were presented on every trial in all three sessions. However, in Sessions 1 and 2 participants in the single-task group performed only one task per trial block, alternating tasks between trial blocks. The dual-task group performed both tasks (two searches) on each trial. Each session consisted of eight trial blocks with nine trials/block and 60 stimuli per trial. During the transfer session the single-task group performed both searches concurrently; the dual-task group just continued its performance of the two tasks. Although the results from the training sessions showed an anticipated initial performance advantage for the single-task group, by the end of the second training session this advantage disappeared, and the dual-task group obtained the same proportion of correct detections and a similar false alarm rate as the single-task group. However, their mean response time was slower. In the transfer condition, the dual-task group detected more targets and responded more quickly than the single-task group.

Comments

Most investigators of multiple-task performance have limited resources. Consequently, they want multiple-task performance to stabilize quickly. The question is how to do that efficiently. That is, what is the best mixture of single- and multiple-task practice to reach multiple-task stability quickly? Is it more efficient to provide some single-task practice before beginning multiple-task performance or should performance on each of the components reach stability first? Based on the studies reviewed in this section, this question cannot be answered with certainty. Tentatively, it appears that any single-task practice beyond one or two trials provides no incremental benefits to subsequent multiple-task performance.

In future research investigators should consider using one of the methods described in this report to identify performance stability. These include examining the inter-trial correlation matrix and using one of the three differential stability techniques. The use of these methods should allow investigators to determine how much component task practice is necessary for the task combination to reach stability quickly. Although use of these measures does require pretesting, they should improve the efficiency of the experimental protocol and produce more interpretable results.

CONTROLLING TASK PRIORITIZATION THROUGH FEEDBACK

Controlling task prioritization is a critical methodological issue in multiple-task studies. Investigators typically attempt to control task prioritization through instructions, such as “pay equal attention to both tasks”. However, the effectiveness of such instructions is uncertain because they are open to interpretation by the participant. Under unequal priority conditions,

the participants' understanding of the instructions is even more in doubt. Consequently, feedback is necessary to assist the participant to conform to the task prioritization instructions.

This section begins with a brief discussion of the types of measures that should be used as feedback, followed by a short section on terminal feedback. The rest of this section is devoted to concurrent feedback, which is often used to control task prioritization under unequal priority conditions. This section concludes with a discussion of some of the problems associated with concurrent feedback.

Which Measures Should Be Presented?

Investigators should consider two issues when selecting performance measures to use as feedback. First, the participant must be able to understand the measure. Correct reaction time, percentage error, and number of responses are easily understood, whereas other measures, such as root mean squared (RMS) error, are not. In such situations, the investigator might calculate two measures: one for feedback that is easily understood and a second for analysis. For example, North et al. (1978) recorded and analyzed RMS (Root Mean Squared) error. However, time on target, a much easier measure to understand, was used as feedback to the participants.

Second, the number of performance measures used as feedback should be limited. Investigators frequently calculate multiple measures for each task of a combination, such as response interval, correct response time, percentage error, and number of omissions. Only those measures that will motivate and assist the participant to follow the task instructions should be provided as feedback

Terminal Feedback

Psychologists usually provide performance feedback after the end of a timed trial or after a fixed number of responses. This type of feedback, terminal feedback, is used commonly in multiple-task studies. Typically, the feedback consists of measures of average performance for each task over a trial and is presented visually. However, some investigators have adopted less traditional methods of providing terminal feedback. For example, participants in a study by Strobach, Liepelt, Pashler, and Schubert (2013) performed two discrete tasks. They were shown their response time after every response under both single- and dual-task conditions. If the response was incorrect, the participants saw only the word "error." Under dual-task conditions, the participants were shown only the fastest of the two response times. In a second example of unusual terminal feedback, Gopher, Brickner, and Navon (1982) gave the

participant verbal feedback at the end of each trial. The authors did not describe how the interactions between the investigator and the participant were controlled.

Concurrent Feedback

As computing power increased and its cost decreased in the 1970's, investigators began experimenting with unequal task priorities. Terminal feedback was found inadequate in such studies. Consequently, investigators began testing different methods for providing concurrent feedback. One method presents concurrent feedback only when performance is poor. For example North and Gopher (1974) used a warning tone to indicate that the response time was too long, or that performance was poor on one of the tasks. This type of discrete feedback provides no information about average performance nor does it help the participant comply with the priority instructions.

Most concurrent feedback, however, is continuous and developed to help the participant comply with the instructions. In one of the early attempts at providing continuous feedback North and Gopher (1976) presented feedback using colored bars that moved up and down. Each bar presented information for one task and was labeled accordingly. Performance scores were scaled so that the higher the bar, the better the performance. North and Gopher decided that, under the initial dual-task conditions, the height of the bar would be based on the mean and standard deviation (SD) of a participant's single-task performance. The distance available on the display was scaled in SD units. The bottom of the display, the zero point of each moving bar (no bar showing), represented -2.5 single-task SDs of the associated task. The midpoint of the display represented the mean single-task performance, and the maximum possible height of the display represented 2.5 single-task SDs above the mean. The height of the performance bar at any moment was essentially a standardized score based on a running average of the participant's recent multiple-task performance and single-task performance. The equation for the momentary height of the performance bars is:

$$\frac{\text{Running average of dual-task performance} - \text{average single-task performance}}{\text{SD single-task performance}}$$

A short horizontal line was drawn across each moving bar display to represent the performance the participant should achieve on that task. These were referred to as "desired performance" lines, and during early dual-task trials, they were positioned based on the participant's single-task performance. In experimental conditions involving changing task priorities, calculation of the height of the moving bars and the desired performance lines was

based on the participant's initial dual-task performance rather than their single-task performance. To ensure that the performance bars moved smoothly, the height of each bar was based on a running average of several seconds of dual-task performance for tracking tasks and a specified number of responses for discrete tasks.

Gopher et al. (1982) used a slightly different feedback indicator. In this experiment, the feedback indicators were horizontal rather than vertical. Both were displayed at the top of the screen with a small distance between them. Only one desired performance line was used, which was placed between the two bars. For experimental conditions involving unequal task priorities, the desired performance line was placed farther away from the end of the higher priority task. The length of the bar was calculated in a manner similar to that of North and Gopher (1976) and reflected the average performance on the task over a short interval.

A more advanced version of the moving bars was used in the Wombat (Roscoe, Corl, & LaRoche, 1997). The Wombat uses tracking as the primary task and three discrete tasks as secondary tasks. Under multiple-task conditions, participants choose which discrete task to perform. Four concurrent feedback indicators are shown on the display screen under multiple-task conditions. The first two indicators are located on the upper left-hand and upper right-hand sides of the screen. One provides feedback about the tracking task; the other, about the secondary task. The indicators are rectangles with dimensions that change. The height of the rectangle represents the participant's recent performance on the task, and the width represents the worth of the task. For the secondary task, the worth is determined by how frequently the participant performs the task. The worth drops if the task is performed too frequently, which forces the participant to perform all three secondary tasks periodically. The performance value (height of the rectangle) is multiplied by the worth of the task (the width) to produce an "effectiveness score."

The third indicator looks similar to one of North and Gopher's (1976) moving bars with some notable differences. The colored portion of the bar represents the total number of points (overall score) the participant has accumulated. The short, horizontal line represents the participant's predicted total score on the Wombat. The fourth indicator is shaped like a triangle and rests on the top of the colored portion of the bar. The point of the triangle reflects the combined "effectiveness score" for the current tasks.

Single-task usage. Moving feedback bars are used almost exclusively under multiple-task conditions. Nevertheless, in at least one instance the bars were used under single-

task conditions. Harris, North, and Owens (1978) initially set the desired performance line at an arbitrary average reaction time of 3 seconds. The height of the moving bar was based on a running average of the last 10 responses. As the average reaction time dropped, the desired performance line was recalibrated to reflect the newest running average. The authors believed that the bars stimulated participants to reach their maximum single-task performance and used them as motivation.

Distraction. Some investigators have suggested that moving feedback indicators might be distracting because the participant must shift their gaze from the task stimuli to the feedback indicator. This is a particular concern when the stimuli for the experimental tasks are visual and the feedback indicators are not within foveal vision. Brickner and Gopher (1981), Gopher, Brickner, and Navon (1982), and Wickens, Tsang, and Pierce (1985) addressed this issue by having participants perform under equal priority conditions with and without the feedback indicators. In all three studies no significant difference in performance was found with or without the feedback indicators.

Comments

One obvious difficulty with using moving performance bars involves tasks with more than one dependent measure, such as speed and accuracy. Which measure should the bars display? At least three methods have been developed to address this problem. In the first method, an investigator manipulates the instructions or a payoff scheme to keep performance on one dependent measure within a restricted range. For example, payoff matrixes that heavily penalize errors are developed and used in conjunction with terminal feedback given at the end of a block of trials. This method might or might not have the desired effect on performance. (See Wang, Proctor, & Pick, 2007 for an example of instructions that were not completely effective). A second method uses a composite measure in lieu of individual dependent variables. One example of this approach is the CRI (Correct Response Interval; see the Analysis Issues section), which ignores incorrect responses. That is, when an incorrect response occurs, the CRI is the time between the preceding correct response and the next correct response including the interval during which an incorrect response was made (Damos & Wickens, 1980). CRIs have greater variance than the correct reaction time (CRT) because some response intervals can be very long, particularly if the participant makes a series of errors. Another example of a composite measure is the response interval (RI), also called the average response interval (ARI), which is the total time in a trial divided by the number of responses (Damos & Wickens, 1980; Gopher & North, 1977). Thus, the RI and the ARI simply ignore errors.

A second more subtle problem concerns how the investigator instantiates the terms “the tasks are equally important” or “give 60% of your attention to the tracking task.” To provide graphic feedback such as the moving performance bars, the investigator must decide what “equal attention” or “equal priority” means in performance units that can be displayed. North and Gopher (1974) used single-task performance as a basis for scaling the moving feedback bars and calculating the position of the desired performance lines for early dual-task trials. However, other approaches to scaling the feedback indicators could have been used. For example, the participants could have been given equal priority instructions for the first one or two dual-task trials. The feedback indicators then could have been based on the participant’s dual-task performance, not on their single-task performance. This approach might instantiate the participant’s understanding of “equal priority” better than using single-task performance as the basis of the feedback.

Investigators should think carefully about how participants will understand the instructions. Does “pay equal attention” mean “make the same number of responses to both tasks when they are both discrete?” Does the participant somehow relate the multiple-task instructions to their single-task performance? What measures best capture the participants understanding of “pay equal attention” or “give 60% of your attention to Task 1?” These questions are just a few that investigators must consider before determining the type of feedback to use.

CONTROLLING RESPONSE STRATEGIES

As noted earlier, the topic of individual differences in response strategies is worthy of investigation in its own right. Nevertheless, allowing participants to adopt their own response strategy when it is not the focus of the study can introduce substantial variance into the data, causing unanticipated and sometimes inexplicable statistical results. More important, unless the response strategy is controlled, the investigator might be unable to study fine-grained “timesharing” behaviors; for example, rapid switching between the tasks or parallel information processing. Several studies (Brüning & Manzey, 2018; Damos, 1984; Damos & Smist, 1982; Damos et al., 1983; Damos & Wickens, 1980; Reissland & Manzey, 2016) (see the Individual Differences in Processing Styles section) demonstrate that under multiple-task conditions, some participants perform one task for long periods before switching to the other. An investigator cannot study fine-grained timesharing behavior if the participants perform one task for an

extended period. The section below describes several methods for controlling the response strategy.

Restricted Response Options

Sverko and his associates (Jerneic, 2007; Sverko et al., 1983, 1994) developed several versions of restricted-response paradigms. The purpose of these paradigms was to limit data collection to periods during which the participant was most likely to be using an information processing strategy of interest to the investigator. In Sverko et al. (1994) and Jerneic (2007) the choice reaction time task combinations were forced-paced with an inter-stimulus interval of 2 seconds. The stimuli for the two tasks were presented simultaneously. If the participant failed to respond to both stimuli correctly within the 2-second interval, the data from that presentation were not included in the analysis. This eliminated the possibility that data from a participant using a massed response strategy would be included in the analysis.

A different restricted-response method was adopted by Sverko et al. (1983) and by Jerneic (2007) for his tracking combinations. The Sverko et al. study employed two forced-paced, discrete task combinations and a tracking/discrete task combination. The participants had to respond “simultaneously” to the stimuli used for the discrete task combination, and the response had to occur within the preset interstimulus interval. For the tracking/discrete task combination, performance on the discrete task was scored only when the tracking cursor was on the target. Sverko et al. adopted this paradigm to control for the tendency to “favor one or the other component task” (p. 153). For the tracking combinations used by Jerneic, performance was scored only when both cursors were on their targets simultaneously. This paradigm increased the likelihood that data would be recorded only when the participants were processing in parallel or switching rapidly between the tasks.

Sverko et al. (1994) noted that the participants made a “sizable” number of omissions but gives no specific numbers. Jerneic (2007) found that even at the highest level of practice, the percentage of simultaneous time on target ranged from 47.9% to 17.8% for the various combinations. For the discrete task combinations the early average correct reaction time exceeded the sum of the single-task averages and then decreased to less than the sum. However, the author does not provide information on the percentage of discrete responses that were either incorrect or completed after the interval. Thus, it is unclear how successfully the participants complied with the task demands.

Two studies discussed earlier by Manzey and his colleagues (Bruning & Manzey, 2018; Reissland & Manzey, 2016) also restricted the participants' response strategy. Because these studies used very similar tasks, only Reissland and Manzey will be discussed. The authors used two versions of a classical task-switching paradigm. In one version, stimuli for only one task were displayed at a time. The participants made three responses to one task before switching to the other. This stimulus presentation method forced the participants to process the tasks serially. In the other version, the participant performed the first task in the same manner but always had a preview of the first stimuli for the second task. Thus, participants could process information for the second task in parallel with information from the first task if they chose to do so. Through a sophisticated set of analyses, Reissland and Manzey were able to identify those who processed the information serially and those who performed some of the processing in parallel.

Comments

Investigators using restricted response protocols are usually concerned with assessing a fine-grained aspect of multiple-task performance. They maintain that the behavior of interest cannot be observed unless the responses are emitted under very controlled conditions that maximize the likelihood that the behavior will occur. That is, rapid task switching cannot be observed unless the participant switches between the tasks. Although this is a valid argument, some participants might not be able to perform acceptably using the restricted paradigm. For example, Bruning and Manzey (2018) excluded 4 of 96 participants because their error rates exceeded 15%. Similarly, Jerneic's (2007) participants were capable of performing the dual-task tracking combination only between 17.8% and 47.9% of the time. Sverko et al.'s (1994) participants made a substantial, but unspecified, number of omissions.

The fact that so much data were excluded in some of these studies can be interpreted in two ways. The first is that the magnitude of the data exclusion calls into question the results and conclusions, limiting their generality. The second is that the inability of some participants to comply supports the notion of individual differences in multiple-task performance. This interpretation raises some interesting questions. For example, were the poor performers in Sverko et al. (1974) and Jerneic (2007) Manzey's blockers and Damos's massed responders? Did the poor performers have intertask switching speeds that were too slow to allow successful multiple-task performance? Such questions must await further research.

ANALYSIS ISSUES

Anyone investigating multiple-task performance is faced with at least two data analysis challenges. The first challenge involves the number of variables. By definition, under multiple-task conditions, the participant must perform two or more tasks concurrently. Consequently, there are at least two dependent variables, and usually more, that reflect the participant's performance. Frequently, additional measures that are unique to multiple-task environments—such as task-switching frequency—are assessed to reflect fine-grained aspects of performance. Analyzing and interpreting a large number of variables can be problematic, and some investigators might want to reduce the number of dependent variables to be analyzed.

The second challenge is accounting for single-task performance effects on multiple-task performance. In much multiple-task performance research, single-task performance is of little interest except as a baseline. The investigator, therefore, must determine how to identify and account for the effects of single-task performance on multiple-task performance. Methods of addressing both of these challenges are discussed below.

The last part of this section deals with control theory measures. These measures have not been extensively explored by psychologists for use in selection despite their widespread use in engineering. They should be given more consideration because they can provide fine-grained measures of tracking performance that allow for new insights into the data.

Reducing the Number of Dependent Variables

Composite measures. Composite measures attempt to incorporate two or more performance measures into a third measure. An example of a composite measure is the correct response interval (CRI). This measure can be calculated two ways. The first ignores any incorrect responses. For example, assume the participant makes a correct response to Stimulus 1, which is immediately replaced with Stimulus 2. The participant then makes an incorrect response to Stimulus 2, which is immediately replaced with Stimulus 3. The participant responds correctly to Stimulus 3. If the investigator were using CRT, the response times to both Stimulus 1 and Stimulus 3 would be used in calculating the mean CRT for the trial, but the response time to Stimulus 2 would be excluded. For the calculation of the CRI, the reaction time to Stimulus 1 would be included in the same way as in the calculation of mean CRT. However, the reaction time to Stimulus 3 would be calculated from the presentation of Stimulus 2, not Stimulus 3, and the reaction time to Stimulus 2 would be excluded from the calculation.

Thus, CRI treats incorrect responses as if they never occurred. Examples of studies using the CRI include Damos and Wickens (1980) and Gopher et al. (1982). The second way to calculate the CRI is to divide the trial length by the number of correct responses made during the trial (see Gopher & North, 1975, for an example).

Another composite measure is the average response interval (ARI), also called the *response interval (RI)*. An example of this occurs in Gopher and North (1974). Under dual-task conditions, the discrete task was forced-paced, and if the participant did not respond within a specified interval, the stimulus was removed and a new one presented. The authors simply divided the length of the trial by the number of both correct and incorrect responses. The authors argue that ARI reflects correct and incorrect responses as well as omissions.

Averaged z scores. Averaging the z scores is a rare method for combining dependent measures in multiple-task studies. Only one example (Rieck et al., 1980) was found. In this study, the authors measured four variables for each task in a discrete task combination. A z score was created for each variable by using data from all of the participants on that task. The four z scores for each discrete task were then averaged to create one score for each combination.

The goal of this approach was to correct for any participant's tendency to favor one task of a combination over another. The participants were instructed to give equal priority to both tasks. This approach might or might not have accomplished this goal. However, it clearly obfuscates any individual differences in within- or between-task speed/accuracy trade-offs. It also eliminates the possibility of any fine-grained analyses of performance.

Correlated common measures. Investigators of multiple-task performance often express multiple-task performance as a proportion of single-task performance (multiple-task performance/single-task performance) or as a percentage of single-task performance (multiple-task performance/single-task performance X 100%). In these cases, the single-task measure is understood as a baseline and the proportion represents the decrement that occurs under timesharing conditions. Proportion scores are intuitively appealing and, again, reduce the number of variable to be analyzed.

A major concern with the use of proportion scores occurs when they have a common denominator and the proportion scores are correlated. This approach is sometimes used to identify consistent individual differences. Typically, a participant receives extensive single-task practice to establish their baseline performance. They then performs a task combination under

different experimental conditions. The investigator uses the participant's baseline performance to create a proportion score for that participant for each experimental condition. The investigator then correlates the proportion scores to identify consistent individual differences in performance decrements.

This paradigm was used by Morgan, Winne, and Dugan (1980), who examined the effect of four work/rest cycles on the performance of the Multi-Task Performance Battery (Chiles, Aluisi, & Adams, 1968). The investigators provided extensive training to establish a baseline score, divided each participants' score during a given work/rest cycle by his baseline score, and subtracted 100%. The resulting score represented a percentage decrement from baseline. The decrement scores were then correlated to test for consistent individual differences.

The problem with correlating proportions that use a common denominator was first noted by Pearson(1897). He demonstrated that the correlation between two proportions using the same denominator is approximately 0.50 even when the original variables that constitute the ratios are uncorrelated. (See also Dunlap, Dietz, & Cortina, 1997.) Bittner (1983) notes that, regardless of the level of significance, correlations below 0.50 are artifacts of the proportion-of-baseline measures and should be disregarded. He also describes problems with the variance and covariance measures of proportion-of-baseline measures.

Eliminating the Effects of Single-Task Performance

Multivariate analysis of covariance. An investigator might study questions that are concerned only with multiple-task performance. One example is "How does the type of concurrent feedback affect multiple-task performance?" Another is "How does response strategy affect inter-task switching speed?" In such situations, the investigator might consider ignoring single-task performance measures and analyzing only the data obtained under multiple-task conditions. However, because the abilities and skills that affect single-task performance also affect multiple-task performance, the investigator should attempt to find a method to remove the effects of the single-task skills and abilities from multiple-task performance rather than ignore them. Multivariate analysis of covariance (MANCOVA) using single-task performance measures as the covariates is one potential approach.

A detailed discussion of the assumptions of MANCOVA is beyond the scope of this paper (See Stevens, 1996, and Tabachnick & Fidell, 1996, for detailed discussions of the assumptions and limitations of the technique.) Nevertheless, two assumptions need to be

addressed. The first is that the covariates are reliable. Covariates that are not reliable can result in increased Type I or Type II errors (Tabachnick & Fidell, 1996, p. 383). Practice typically increases reliability of both speed and accuracy measures. However, for practical reasons, most investigators have to limit the amount of single-task practice given on each component task. The investigator should use the differential stability methods described in the Practice section to determine how much practice is enough for each component task.

The second assumption is that the relation between each pair of covariates and between each multiple-task dependent variable—covariate pair is linear. Given that the speed/accuracy trade-off can change with practice, the investigator again will have to examine carefully the effect of practice on the pairwise relation between the covariates and between the dependent measures and the covariates. If a relation is clearly nonlinear, it might be possible to transform the covariate or use a higher-order power of the covariate to make the relation linear. In extreme cases, the covariate might have to be dropped (See Cohen & Cohen, 1975 for a detailed discussion of nonlinear covariates.)

Partialling single task. For investigators identifying a general timesharing ability, a method for eliminating the effects of single-task performance on multiple-task performance was developed by Bittner and Damos(1986). The first step in this process requires the investigator to use partial correlation analysis to remove the variance attributed to single-task measures from multiple-task measures. More specifically, all of the single-task measures from a given task are partialled out of all of the other single-task measures and out of all of the multiple-task measures before the data are analyzed. Although some might question why the variance attributed to a given single-task performance measure should be partialled out of all of the other single-task performance measures and out of combinations that do not include the task, this step is necessary for three reasons. First, single-task measures are often correlated with each other. Second, a performance measure from a given task often correlates with all dual-task measures, even when the task is not part of a combination. Thus, removing all of the single-task variance that contributes to multiple-task performance makes identifying unique timesharing factors easier. Third, partialling out all of the variance attributed to single-task performance allows the maximum variance attributable to timesharing factors to be identified. During the second step of the process, the multiple-task partial correlations are submitted to an iterative principal factor analysis with varimax rotation.

Bittner and Damos (1986) reanalyzed the Wickens et al. (1981) data using the method described above and compared their results with those of Ackerman et al. (1984).¹ Although both analyses found a four-factor solution, the results were substantially different; Bittner and Damos found no evidence of a general timesharing factor. This difference might be attributed to the partialling of single-task variance from the performance measures but also to the specific type of analysis used.

Ackerman et al. used an orthogonal Procrustes procedure, which is problematic for several reasons. One of these relates to the target structure, which should be based on a theory or derived from the results of previous studies (McCrae, Zonderman, Costa, Bond, & Paunonen, 1996, p. 558). The source of the target structure in the Ackerman et al. reanalysis is unclear. It might be based on a model, briefly described on page 72, that assumes “error free data with perfect reliability” or on the Wickens et al. data. If the target structure was based on the Wickens et al. data, then the results are subject to the criticism of Fokkema and Greiff (2017), who demonstrated “deceivingly optimistic (p. 400)” model fits when the results of a principal component analysis were used to guide a confirmatory factor analysis conducted on the same data. Another problem concerns the Procrustes procedure itself, which has been criticized for its subjectivity (Horn, 1967; Humphreys, Ilgen, McGrath, & Montanelli, 1969; McCrae et al., 1996). However, perhaps most important, the Procrustes procedure can result, as demonstrated by Horn (1967), in “meaningful” factors when the data are random variables.

Bittner and Damos (1986) also reanalyzed the Sverko 1977 data. Sverko used a principal component analysis and obtained both a four- and a five-factor solution. Neither solution showed evidence of a general timesharing factor. Each of the factors was defined by an experimental task, with the single-task performance measure loading heavily on its associated factor. The reanalysis obtained a four-factor solution, also with no evidence of a general timesharing factor. As with Sverko’s results, each factor was identified with one of the four experimental tasks. Because there is no a priori reason for a dual-task factor to be defined by an experimental task, the results might reflect some aspect of the methodology or a set of specific timesharing factors.

¹ Only the intercorrelations of the single-task measures are given in Wickens et al. (1981). The correlations between the single- and dual-task measures and among the dual-task measures were provided by Wickens.

Control Theory Measures

Control theory measures provide a level of insight into tracking behavior that is not possible with measures commonly used by psychologists, such as time on target. A description of control theory is beyond the scope of this report, but good introductions are available in Damos and Wickens (1980) and Wickens and Gopher (1977). As described in Damos and Wickens, under dual-task conditions, control theory measures can detect periods of serial processing, non-independent processing (“cross talk” between the two tasks), and parallel processing. Changes in the amount of any of these three types of processing can be detected as a function of practice or different experimental conditions.

The cost of software programs that compute control theory measures has dropped dramatically over the last 20 years. Having sufficient computing power to obtain the data and perform the analyses is no longer an issue. Control theory has a strong theoretical basis and is widely used in several branches of engineering. It should be a welcome addition to the analysis techniques used by psychologists.

Comments

Investigators studying multiple-task performance often must deal with two major assessment issues: decreasing the number of dependent measures and accounting for the contribution of single-task performance to multiple-task performance. Several methods for dealing with these topics were presented in this chapter. Some of these, such as partialling single-task performance from multiple-task performance, are useful in answering only certain questions. Others, such as composite measures, might be useful in a variety of studies.

One major assessment issue, performance trade-offs, was not discussed in this report. Individual differences in both intra-task and inter-task trade-offs can have a profound effect on the results of the statistical analysis. Under single-task conditions, a participant might adopt one speed/accuracy trade-off for a given task but change the trade-off with practice or with different experimental conditions. Other participants may adopt very different speed/accuracy trade-offs for the same task. When the participant performs the task under timesharing conditions, they may adopt a different speed/accuracy trade-off or may trade-off speed on one task of the combination for accuracy on another. This trade-off may change with practice and with changes in the experimental conditions.

Such changes may result in heteroscedasticity of variance and changes in the covariance structure with practice and experimental conditions. Two approaches discussed earlier may be used to restrict the intra- and inter-task trade-offs. The first, and perhaps the most effective, is through the use of both concurrent and terminal feedback. Both types of feedback, but particularly concurrent feedback, should be effective in reducing between-participant differences in trade-off strategies. Clear, precise instructions are also necessary to control trade-offs. The second approach is through restricting the participants' response strategy. This approach has drawbacks, such as the elimination of a substantial proportion of responses, which were discussed previously.

If both of these methods fail to control the intra- and inter-task trade-offs, the investigator may consider using the composite measures discussed earlier. Proponents of these measures maintain they take into account correct responses, incorrect responses, and omissions. However, no studies comparing these measures (for example, CRI versus CRT and percentage error) were found. Further, little is known about the statistical characteristics of these measures. Consequently, the investigator should proceed cautiously with these measures.

Control theory has been used in human factors work for more than 40 years. However, it has not been adopted in other areas of psychology, perhaps because of its apparent mathematical complexity. Control theory measures appear never to have been used for assessment of pilot training aptitude. Its analyses can provide a completely different understanding of tracking performance under both single- and multiple-task conditions that might prove helpful in selecting and classifying pilot training applicants.

CONCLUSIONS

The purpose of this report was to provide an overview of the areas of multiple-task performance that could be useful for personnel selection, particularly pilot training selection, and to review topics that have caused problems assessing multiple-task performance.

Understanding of multiple-task performance has not advanced quickly over the last 40 years because of the challenges inherent in dealing with correlated, multiple dependent variables that reflect intra- and inter-task trade-offs that change with practice and differ between individuals. Investigators have attempted to address this problem in two ways. First, they have employed a variety of statistical techniques, many of which ignore some aspects of performance. Second, they have used highly constrained methodologies that allow the study participant no freedom to

organize their responses. These methodologies effectively prohibit the participant from making an error or omitting a response. Neither of these methods deals effectively with the problem of trade-offs.

It might be possible to address the trade-off issue by investigating other branches of psychology and science. Most investigators working in multiple-task performance were trained in either experimental psychology or human factors. Those in areas such as econometrics, biology, and chemical engineering might have statistical approaches to complex, multivariate problems that could be applied to multiple-task performance but have not yet been investigated. Another potentially beneficial area is human performance models. Sophisticated human performance models had to instantiate intra- and intertask trade-offs. A. C. Bittner (personal communication, October 18, 2019) indicated that non-linear inter-task trade-offs were included in the Human Operator Simulator (HOS) model and in other human performance models that were developed at approximately the same time as HOS. Newer models, such as the Strategic Task Overload Model (STOM) (Wickens & Gutzwiller, 2017) and the MicroSaint models, also are worth investigating to determine how trade-offs are instantiated.

What is the status of individual differences in multiple-task performance? As documented in this report, relatively little research has focused on individual differences. The research that has been conducted has been concentrated on two major topics: the identification of a general timesharing ability and response strategies. Research conducted over approximately 120 years shows little evidence of a general timesharing ability. Results from a few studies (Braune & Wickens, 1986; Jerneic, 2007) suggest the existence of some specific timesharing abilities, but no research appears to have been conducted on either the structure of timesharing abilities or on the identification of specific abilities.

Several studies (Brüning & Manzey, 2018; Damos, 1984; Damos & Smist, 1982; Damos et al., 1983; Damos & Wickens, 1980; Reissland & Manzey, 2016) have identified individual differences in multiple-task performance. These results are encouragingly similar given that Manzey's studies were conducted in a different country approximately 30 years after those by Damos'. Nevertheless, there is little evidence that the response pattern approach to individual differences is being pursued by other investigators.

Prospective memory was investigated as a potential new source of individual differences in multiple-task performance. However, database searches revealed no evidence of studies reporting test-retest reliabilities, and gender and ethnic differences do not seem to have been

investigated. Additionally, prospective memory has not yet been located in a specific brain region as have other memory systems. The applied research that has been conducted leaves the impression that prospective memory is more of a paradigm than a physiologically-based source of individual differences. Although prospective memory is not currently a viable attribute for selection, it is being studied by investigators in several fields. At some point in the future, it might reveal a new source of individual differences in timesharing.

Recommendations

Five recommendations are made to support future USAF efforts in developing multiple-task tests to improve pilot selection and/or classification.

1. Consider adding a discrete task combination to the Test of Basic Aviation Skills (TBAS). The Braune and Wickens (1986) and Jerneic (2007) studies suggest that timesharing abilities may be specific to major categories of tasks, such as tracking. TBAS currently has several combinations that involve tracking, but it has no discrete task combinations. The addition of a discrete task combination might enable the assessment of different timesharing abilities. It also opens the possibility of assessing more fine-grained behaviors, such as intertask switching speed.
2. Consider performing control theory analyses on the TBAS tracking data. The data analysis programs have dropped substantially in price, and computing power is no longer a consideration. The control theory analyses will produce new measures that provide a fine-grained analysis of tracking performance.
3. Review new statistical methods in psychology and other areas of science that can be used to deal with the intra- and inter-task trade-offs that differ between participants and over time. The traditional statistical techniques are simply inadequate for analyzing these types of trade-offs. Similarly, the problem of analyzing correlated dependent measures obtained under both single- and dual-task performance needs to be revisited.
4. Review the literature and documentation on human performance models, such as HOS, as well as more recent computational and human factors models, to determine how multiple performance measures are incorporated into the model. The models of vehicle drivers, especially those using optimal control theory, also should be reviewed.

5. Extend the research conducted by Manzey and his colleagues (Brüning & Manzey, 2018; Reissland & Manzey, 2016) and Damos and her colleagues (Damos, 1984; Damos & Smist, 1982; Damos et al., 1983; Damos & Wickens, 1980). The performance of different response strategy groups needs to be examined in more complex multiple-task combinations, particularly those including tracking.

REFERENCES

- Ackerman, P. L., Schneider, W., & Wickens, C. D. (1984). Deciding the existence of a time-sharing ability; a combined methodological and theoretical approach. *Human Factors*, 26, 71-82.
- Bartram, D., & Baxter, P. (1996). Validation of the Cathay Pacific Airways Pilot selection Program. *International Journal of Aviation Psychology*, 6(2), 149-170.
- Bittner, A. C., Jr. (1979). Statistical tests for differential stability. In *Proceedings of the Human Factors Society Annual Meeting* (pp. 541-545). Washington, DC: Human Factors Society.
- Bittner, A. C., Jr. (1983). Use of proportion-of-baseline measures in stress research. In G. Salvendy & M. J. Smith (Eds.), *Machine pacing and occupation stress* (pp. 177-183). Florence, KY: Taylor & Francis.
- Bittner, A. C., Jr., & Damos, D. L. (1986). *Demonstration of a statistical method for isolating timesharing components*. (NBDL-86R001). New Orleans, LA: Naval Biodynamics Laboratory
- Bittner, A. C., Jr., & Damos, D. L. (1988). Statistical isolation of timesharing components: demonstration and interpretation. In F. Aghazadeh (Ed.), *Trends in Ergonomics/Human Factors V* (pp. 21-27). NY: Elsevier.
- Bradley, J. V. (1969). Practice to an asymptote? *Journal of Motor Behavior*, 1(4), 285-296.
- Braune, R., & Wickens, C. D. (1986). Time-sharing revisited: Test of a componential model for the assessment of individual differences. *Ergonomics*, 29, 1399-1414.
- Brickner, M., & Gopher, D. (1981). *Improving time-sharing performance by enhancing voluntary control on processing resources*. (TR-82-0633). Bolling AFB, DC: Air Force Office of Scientific Research.
- Brookings, J. B. (1990). A confirmatory factor analytic study of time-sharing performance and cognitive abilities. *Intelligence*, 14, 43-59.
- Brookings, J. B., & Damos, D. L. (1991). Individual differences in multiple-task performance. In D. L. Damos (Ed.), *Multiple-Task Performance* (pp. 363-386). London: Taylor & Francis.
- Brüning, J., & Manzey, D. (2018). Flexibility of individual multitasking strategies in task-switching with preview: Are preferences for serial versus overlapping task processing dependent on between-task conflict? *Psychological Research*, 82, 92-108.
- Carretta, T. R. (1990). *Cross-validation of experimental USAF pilot training performance models*. (AFHRL-TR-89-68). Brooks AFB, TX: Air Force Human Resources Laboratory
- Carretta, T. R. (2011). Pilot candidate selection method: Still an effective predictor of US Air Force pilot training performance. *Aviation Psychology and Applied Human Factors*, 1(1), 3-8.
- Carretta, T. R. (2013). Predictive validity of pilot selection instruments for remotely piloted aircraft training outcomes. *Aviation, Space and Environmental Medicine*, 84(1), 47-53.
- Carretta, T. R., Zelenski, W. E., & Ree, M. J. (2000). Basic Attributes Test (BAT) Retest Performance. *Military Psychology*, 12(3), 221-232.
- Carroll, J. B. (1993). *Human cognitive abilities A survey of factor-analytic studies*. NY: Cambridge University Press.
- Chiles, W. D., Aluisi, E. A., & Adams, O. S. (1968). Work schedules and performance during confinement. *Human Factors*, 10, 143-196.
- Cohen, J., & Cohen, P. (1975). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Hillsdale, NJ: Erlbaum.
- Damos, D. L. (1984). Individual differences in multiple-task performance and subjective estimates of workload. *Perceptual and Motor Skills*, 59, 567-580.
- Damos, D. L. (1986). *The effect of the amount of single-task practice on the performance of discrete task combinations*. (Report No. 916745-TE). Tempe, AZ: Department of Psychology, Arizona State University.

- Damos, D. L. (1987). *The effect of the amount of single-task practice on dual-task tracking*. (Report No. 916745-TE2). Los Angeles: Dept. of Human Factors, University of Southern California.
- Damos, D. L. (1991). Dual-task methodology: Some common problems. In D. L. Damos (Ed.), *Multiple-Task Performance* (pp. 101-119). London: Taylor & Francis.
- Damos, D. L., Bittner, A. C., Jr., Kennedy, R. S., & Harbeson, M. M. (1981). Effects of extended practice on dual-task tracking performance. *Human Factors*, 23(5), 628-631.
- Damos, D. L., & Lintern, G. (1981). A comparison of single- and dual-tasks measures to predict simulator performance of beginning student pilots. *Ergonomics*, 24(9), 673-684.
- Damos, D. L., & Smist, T. E. (1982). Individual differences in multi-task response strategies. *Aviation, Space, and Environmental Medicine*, 53(12), 1177-1181.
- Damos, D. L., Smist, T. E., & Bittner, A. C., Jr. (1983). Individual differences in multiple-task performance as a function of response strategy. *Human Factors*, 25(2), 215-226.
- Damos, D. L., & Wickens, C. D. (1980). The identification and transfer of timesharing skills. *Acta Psychologica*, 46, 15-39.
- Detweiller, M. C., & Lundy, D. H. (1995). Effects of single- and dual-task practice on acquiring dual-task skill. *Human Factors*, 37(1), 193-211.
- Dismukes, R. K. (2012). Prospective memory in workplace and everyday situations. *Current Directions in Psychological Science*, 21(4), 215-220.
- Dismukes, R. K., Kochan, J. A., & Goldsmith, T. E. (2018). Flight crew errors in challenging and stressful situations. *Aviation Psychology and Applied Human Factors*, 8(1), 35-46.
- Doat, B. (1995). Need of new development in Air France selection. In N. Johnston, R. Fuller, & N. McDonald (Eds.), *Aviation Psychology: Training and Selection* (Vol. 2, pp. 182-187). Aldershot: Avebury.
- Dunlap, W. P., Dietz, J., & Cortina, J. M. (1997). The suprious correlation of ratios that have common variables: A Monte Carlo examination of Pearson's formula. *The Journal of General Psychology*, 124(2), 182-193.
- Einstein, G. O., McDaniel, M. A., Williford, C. L., Pagan, J. L., & Dismukes, R. K. (2003). Forgetting of intentions in demanding situations is rapid. *Journal of Experimental Psychology: Applied*, 9(3), 147-162.
- Eysenck, H. J., & Eysenck, S. B. G. (1975). *Manual of the Eysenck Personality Questionnaire*. London: Hodder & Stoughton.
- Fitts, P. M. (1962). Factors in complex skill learning. In R. Glaser (Ed.), *Training Research and Education* (pp. 177-197). Pittsburgh: University of Pittsburgh Press.
- Fogarty, G. (1987). Timesharing in relation to broad ability domains. *Intelligence*, 11, 207-231.
- Fokkema, M., & GRIEFF, S. (2017). How performing PCA and CFA on the same data equals trouble. *European Journal of Psychological Assessment*, 33(6), 399-402.
- Gopher, D., Brickner, A., & Navon, D. (1982). Different difficulty manipulations interact differently with task emphasis: Evidence for multiple resources. *Journal of Experimental Psychology: Human Perception and Performance*, 8, 146-157.
- Gopher, D., & North, R. A. (1974). The measurement of attention capacity through concurrent task performance with individual difficulty levels and shifting priorities. In *Proceedings of the 18th Annual Meeting of the Human Factors Society*. Washington, DC: Human Factors Society.
- Gopher, D., & North, R. A. (1975). Determinants of performance improvement in training under time-sharing conditions. In *Proceedings of the 19th Annual Meeting of the Human Factors Society* (pp. 179-185). Washington, DC: Human Factors Society.
- Gopher, D., & North, R. A. (1977). Manipulating the conditions of training in time-sharing performance. *Human Factors*, 19, 583-594.
- Grundgeiger, T., Sanderson, P. M., & Dismukes, R. K. (2014). Prospective memory in complex sociotechnical systems. *Zeitschrift für Psychologie*, 222(2), 100-109.

- Harris, R. L., Sr., & Spady, A. A., Jr. (1985). Visual Scanning behavior. In *Proceedings of the IEEE 1985 National Aerospace and Electronics Conference* (Vol. 2, pp. 1032-1037). Dayton, OH: IEEE.
- Harris, S. D., North, R. A., & Owens, J. M. (1978). A system for the assessment of human performance in concurrent verbal and manual control tasks. *Behavior Research Methods, Instruments, & Computers*, 10(2), 329-333.
- Hawkins, H. L., Rodriguez, E., & Reicher, G. M. (1979). *Is timesharing a general ability?* (Technical Report No. 3). Eugene, Oregon: University of Oregon, Center for Cognitive and Perceptual Research.
- Hoermann, H.-J., Noser, P., & Stelling, D. (2018, September). *A psychometrician's dream: Test validities in an unrestricted sample of future Swiss air force pilots*. Paper presented at the 33rd EAAP Conference, Dubrovnik, Croatia.
- Horn, J. L. (1967). On subjectivity in factor analysis. *Educational and Psychological Measurement*, 27, 811-820.
- Huelmann, G., & Oubaid, V. (2004). Computer assisted testing (CAT) in aviation psychology. In K.-M. Goeters (Ed.), *Aviation psychology: Practice and research* (pp. 123-134). Farnham: Ashgate.
- Humphreys, L. G., Ilgren, D., McGrath, D., & Montanelli, R. (1969). Capitalization on chance in rotation of factors. *Educational and Psychological Measurement*, 29, 259-271.
- Jennings, A. E., & Chiles, W. D. (1977). An investigation of time-sharing ability as a factor in complex performance. *Human Factors*, 19, 535-548.
- Jerneic, Z. (2007). Factor structure of time-sharing performance at different levels of practice. *Review of Psychology*, 14(2), 115-138.
- Jerneic, Z., & Sverko, B. (1994). Time-sharing factors and their relation to cognitive abilities and personality traits. *Personality and Individual Differences*, 16(2), 297-308.
- Jones, M. B. (1970a). Rate and terminal processes in skill acquisition. *American Journal of Psychology*, 83, 222-236.
- Jones, M. B. (1970b). A two process theory of individual differences in motor learning. *Psychological Review*, 77(4), 353-360.
- Jones, M. B., Kennedy, R. S., & Bittner, A. C., Jr. (1981). A video game for performance testing. *American Journal of Psychology*, 94(1), 143-152.
- LaRoche, J. (2011, April). *CQFA--Chicoutimi College Aviation Program and Selection Process*. Paper presented at the World Aviation Training Summit, Orlando, FL.
- Lawley, D. N. (1963). On testing a set of correlation coefficients for equality. *Annals of Mathematical Statistics*, 34, 149-151.
- Mashburn, N. C. (1934). Mashburn automatic serial action apparatus for detecting flying aptitude. *Journal of Aviation Medicine*, 5, 155-160.
- McCrae, R. R., Zonderman, A. B., Costa, P. T., Jr., Bond, M. H., & Paunonen, S. V. (1996). Evaluating replicability of factors in the Revised NEO Personality Inventory: Confirmatory factor analysis versus Procrustes rotation. *Journal of Personality and Social Psychology*, 70(3), 552-566.
- McQueen, E. N. (1917). The distribution of attention. *British Journal of Psychology, II Monograph Supplements* 5.
- Melton, A. W. (1947). *Apparatus Tests Report No. 4*. Washington, D.C.: U.S. Government Printing Office.
- Morgan, B. B., Jr., Winne, P. S., & Dugan, J. (1980). The range and consistency of individual differences in continuous work. *Human Factors*, 22(3), 331-340.
- North, R. A., & Gopher, D. (1974). Basic attention measures as predictors of success in flight training. In *Proceedings of the 18th Annual Meeting of the Human Factors Society* (pp. 50-56). Washington, DC: Human Factors Society.

- North, R. A., & Gopher, D. (1976). Measures of attention as predictors of flight performance. *Human Factors*, 18(1), 1-14.
- North, R. A., Harris, S. D., & Owens, J. M. (1978). *Test-retest reliability of individual differences in dual-task performance*. (Report No. NAMRL-1248). Pensacola, FL: Naval Aerospace Medical Research Laboratory.
- Nowinski, J. L., Holbrook, J. B., & Dismukes, R. K. (2003). Human memory and cockpit operations: An ASRS study. In *12th International Symposium on Aviation Psychology* (Vol. 2, pp. 888-893). Dayton, OH: Wright State University.
- Pearson, K. (1897). Mathematical contributions to the theory of evolution.—On a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the Royal Society of London*, 60, 489-498.
- Reissland, J., & Manzey, D. (2016). Serial or overlapping processing in multitasking as individual preference: Effects of stimulus preview on task switching and concurrent dual-task performance. *Acta Psychologica*, 168, 27-40.
- Rieck, A. M., Ogden, G. D., & Anderson, N. S. (1980). An investigation of varying amounts of component task practice on dual task performance. *Human Factors*, 22, 373-384.
- Roscoe, S. N., Corl, L., & LaRoche, J. (1997). *Predicting Human Performance*. Pierrefonds QC: Helio Press, Inc.
- Sharp, E. (1899). Individual psychology: A study in psychological method. *The American Journal of Psychology*, 10, 329-391.
- Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review*, 84, 127-190.
- Spady, A. A., Jr., Jones, D. H., Coates, G. D., & Kirby, R. H. (1982). The effectiveness of using real-time eye scanning information for pilot training. In *Proceedings of the 26th Annual Meeting of the Human Factors Society* (pp. 1014-1017). Washington, DC: Human Factors Society.
- Strobach, T., Liepelt, R., Pashler, H., & Schubert, T. (2013). Effects of extensive dual-task practice on processing stages in simultaneous choice tasks. *Attention Perception and Psychophysics*, 75, 900-920.
- Sverko, B. (1977). Individual differences in time sharing performance. *Acta Instituti Psychologici Universitatis Zagrebiensis*, 80, 17-30.
- Sverko, B., Jerneic, Z., & Kulenovic, A. (1983). A contribution to the investigation of timesharing ability. *Ergonomics*, 26, 151-160.
- Sverko, B., Maslic-Sersic, D., Jerneic, Z., & Gurdulic-Sverko, A. (1994). Time sharing performance of choice-response tasks: factor structure and age differences. *Review of Psychology*, 1(1), 5-10.
- Tabachnick, B. G., & Fidell, L. S. (1996). *Using multivariate statistics*. New York: HarperCollins.
- Trankell, A. (1959). The psychologist as an instrument of prediction. *Journal of Applied Psychology*, 45(3), 170-175.
- Van Benthem, K. D., Herdman, C. M., Tolton, R. G., & LeFevre, J. (2015). Prospective memory failures in aviation: Effects of cue salience, workload, and individual differences. *Aerospace Medicine and Human Performance*, 86(4), 366-273.
- Wang, D.-Y. D., Proctor, R. W., & Pick, D. F. (2007). Acquisition and transfer of attention allocation strategies in a multiple-task work environment. *Human Factors*, 49(6), 995-1004.
- West, R., Scalaro, A. J., & Bailey, K. (2001). When goals collide: The interaction between prospective memory and task. *Canadian Journal of Experimental Psychology*, 65(1), 38-47.
- Wickens, C. D. (1980). The structure of attentional resources. In J. Requin (Ed.), *Attention and Performance VIII* (pp. 239-259). Hillsdale, NJ: Erlbaum.

- Wickens, C. D. (1992). *Engineering Psychology and Human Performance* (2nd ed.). NY: HarperCollins
- Wickens, C. D., & Gopher, D. (1977). Control theory measures of tracking as indices of attention allocation strategies. *Human Factors*, 19(4), 349-365.
- Wickens, C. D., & Gutzwiller, R. S. (2017). The status of the Strategic Task Overload Model (STOM) for predicting multi-task management. In *Proceedings of the Human Factors and Ergonomics Society 2017 Annual Meeting* (pp. 757-761). Washington, D.C.: Human Factors and Ergonomics Society
- Wickens, C. D., Mountford, S. J., & Schreiner, W. (1981). Multiple resources, task-hemispheric integrity, and individual differences in time-sharing. *Human Factors*, 23, 211-230.
- Wickens, C. D., Tsang, P. S., & Pierce, B. (1985). The dynamics of resource allocation. In W. Rouse (Ed.), *Advances in Man-Machine Systems Research* (Vol. 2, pp. 1-49). Greenwich: JAI Press.
- Wightman, D. C., & Lintern, G. (1985). Part-task training for tracking and manual control. *Human Factors*, 27, 267-283.
- Zuber, S., Kliegel, M., & Ihle, A. (2016). An individual difference perspective on focal versus nonfocal prospective memory. *Memory and Cognition*, 44, 1192-1203.