Available online at www.sciencedirect.com
SCIENCE @ DIRECT®
ELSEVIER
Intelligence 31 (2003) 567–587
INTELLIGENCE

A review of the worst performance rule: Evidence, theory, and alternative hypotheses

Thomas R. Coyle*

Department of Psychology, University of Texas at San Antonio, San Antonio, TX 78249-0652, USA

Received 2 May 2002; received in revised form 31 January 2003; accepted 7 April 2003

Abstract

This paper reviews evidence, theory, and alternative hypotheses for the worst performance rule (WPR), which states that on multitrial cognitive tasks, worst performance trials predict general intelligence (g) better than best performance trials. A review of the relevant evidence indicates that the WPR has been found for a variety of participants, tasks, and measures. A review of relevant theories reveals that the WPR appears to be related to cognitive factors (e.g., lapses in working memory) as well as biological factors (e.g., individual differences in neural oscillations). A review of alternative hypotheses shows that the WPR cannot be attributed to statistical or data artifacts such as outliers, unreliable measurement, or variance compression. The preponderance of evidence supports the hypothesis that the WPR holds for cognitive tasks high in g saturation but not for cognitive tasks low in g saturation. The paper ends with a call for research on the causes of the WPR and for research on the correlates of best performance.

Keywords: Worst performance rule; General intelligence; Worst performance trials; Best performance trials

1. Introduction

Suppose a group of participants receives 30 trials of a choice reaction time task and is given an IQ test. Further suppose that each participant's reaction times (RTs) are ranked from

* Tel.: +1-210-458-7407; fax: +1-210-458-5728.
E-mail address: tcoyle@utsa.edu (T.R. Coyle).

0160-2896/03/$ – see front matter
doi:10.1016/S0160-2896(03)00054-0

fastest to slowest and correlated with IQ. Which of the ranked RTs should be the best predictor of participants' IQs, the slowest RTs or the fastest RTs? On the basis of prior research, the likely answer is that the slowest RTs should be the best predictor of participants' IQs. Such a finding would be consistent with the worst performance rule (WPR), which states that on multitrial cognitive tasks, worst performance trials are more indicative of general intelligence (*g*) than are best performance trials (Larson & Alderton, 1990).

This paper reviews research and theory on the WPR. It is divided into five major sections. Section 2 reviews extant research on the WPR. Section 3 discusses the broader implications of the WPR including its relation to classical measurement theory and to psychometric *g*. Section 4 reviews possible causes of the WPR including lapses in attention and working memory as well as disruptions in basic brain processes. Section 5 evaluates the claim that the WPR can be attributed to data artifacts such as outliers, variance compression, and measurement unreliability. Section 6 reviews issues for future research including tests of the physiological correlates of the WPR and the heritability of worst and best performances.

The review will show that the WPR (a) is obtained across a variety of tasks, measures, and participants; (b) challenges the assumption that extreme scores are loaded with error; (c) appears related to cognitive and neural correlates of psychometric *g*; (d) cannot readily be attributed to statistical or data artifacts; and (e) raises new questions about psychometric *g* and the correlates of best performance.

2. Evidence for the WPR

A test of the WPR is called a worst performance analysis. A worst performance analysis involves three steps. First, participants with different levels of intelligence receive repeated trials of a cognitive task. Second, worst and best performance scores are identified within participants. Finally, worst and best performances are correlated with individual differences in *g*-loaded measures such as IQ. The WPR is indicated (a) when worst performance predicts *g*-loaded measures better than best performance, or (b) when mean levels of worst performance discriminate among IQ groups better than mean levels of best performance.

Eight studies have provided data for a worst performance analysis. Of these, seven studies (Baumeister & Kellas, 1968; Coyle, 2001, 2003; Diascro & Brody, 1993; Jensen, 1982; Kranzler, 1992a; Larson & Alderton, 1990) reported evidence supporting the WPR. One study (Salthouse, 1998) reported no support for the WPR. The eight studies will be reviewed chronologically with the exception of the nonsupportive study, which is reviewed last.

2.1. Early evidence for the WPR

Baumeister and Kellas (1968) were the first to report data consistent with the WPR. In their study, six retarded adults (mean age = 21.4 years, mean IQ = 62) and six university graduate students (mean age = 25.7 years, mean IQ not reported) received 600 trials (300 trials per day over 2 days) of an auditory RT task. Participants were instructed to depress a telegraph key and, upon hearing an auditory tone, to release the key as quickly as possible. RT was the time taken to release the telegraph key after the tone was presented.

Frequency distributions of RTs for the retarded adults and for the graduate students were reported (Baumeister & Kellas, 1968, Fig. 1, p. 716). Whereas the 20 fastest RTs for the two groups were quite similar (RTs = 191 and 113 ms, for the retarded and normal adults, respectively), the slowest RTs were very different (RTs ~ 640 and 380 ms, for the retarded and normal adults, respectively; interpolated from Fig. 1, p. 716). On the basis of these findings, Baumeister and Kellas argued that there were two populations of RT trials: the fastest RT trials, which were similar for all participants regardless of mental ability; and the slowest RT trials, which discriminated between retarded and normal participants. This claim anticipates the WPR, which states that worst performance trials reveal more about intelligence than best performance trials.

Jensen (1982, p. 115–117) replicated and extended Baumeister and Kellas's (1968) findings. Jensen surmised from Baumeister and Kellas's study that if retarded and normal adults had their RTs ranked from fastest to slowest, group differences in RT would be most pronounced for the slowest RTs. He also surmised that the group differences would be exaggerated on more complex RT tasks, which have higher *g* loadings than less complex RT tasks.

To test these hypotheses, Jensen (1982) administered 15 trials of a simple RT task (one response possible) and 15 trials of a choice RT task (eight responses possible) to 46 mildly mentally retarded adults (mean IQ = 70) and 50 bright normal young adults (mean IQ = 120). Participants' RT data were subjected to a worst performance analysis. First, each participant's slowest RT was eliminated to minimize the effect of outliers. Second, each participant's RTs were ranked from fastest to slowest. Finally, the median RT at each rank was computed and the data were plotted by group (retarded or normal adults) and RT task (simple or choice RT).

Jensen's (1982) results indicated that RTs were generally slower for retarded adults than for normal adults. However, the magnitude of this group difference varied with RT task and RT rank. For simple RT, the group difference in RT was least pronounced for the fastest RTs (111 ms) and most pronounced for the slowest RTs (370 ms, interpolated from Jensen, 1982, Fig. 18, p. 116). For choice RT, the same pattern was observed but the group difference was particularly exaggerated for the slowest RTs (RT differences = 142 and 1100 ms, interpolated from Jensen, 1982, Fig. 19, p. 116). On the basis of these findings, Jensen argued that group differences in RT among participants differing in mental ability are amplified by increasing task complexity, which is related to a task's *g* loading.

2.2. Larson and Alderton (1990)

Larson and Alderton (1990) were the first to explicitly test and label the WPR. They administered an 82 trial RT task (the arrows test) to a large sample of Navy recruits ($N = 332$, mean age = 19.8 years). The RT data were subjected to a worst performance analysis. First, each participant's fastest and slowest RTs were deleted and the remaining 80 RTs were ranked from fastest to slowest. Second, the ranked RTs were divided into 16 consecutive RT bands with five RTs per band. Third, Spearman rank-order correlations (r_s) were computed between the median RT of each band and measures of general intelligence, working memory, and clerical speed. General intelligence (*g*) was a composite measure of several subtests of the Armed Services Vocational Aptitude Battery (ASVAB) and of the Raven's Progressive Matrices. Working memory was a composite measure of tests that required participants to simultaneously hold, manipulate, and store information in memory. Clerical speed was a composite measure of two ASVAB subtests (Numerical Operations and Coding Speed) that required participants to solve simple math problems or associate numbers and words as quickly as possible.

Larson and Alderton (1990) found that the correlations with general intelligence increased linearly from the fastest ($r_s = -.20$) to the slowest RT band ($r_s = -.37$). Similar results were obtained for the correlations with working memory ($r_s = -.21$ and $-.36$, for the fastest and slowest RT bands, respectively), a construct theoretically related to general intelligence (Larson & Saccuzzo, 1989). Although the correlations with clerical speed (a construct not theoretically related to general intelligence) generally increased from the fastest to the slowest RT bands, the increase was less consistent and smaller in magnitude ($r_s = -.10$ and $-.17$, for the fastest and slowest RT bands, respectively). On the basis of these results, Larson and Alderton argued that, "The worst RT trials reveal more about intelligence than do other portions of the RT distribution" (p. 310).

2.3. Kranzler (1992a)

Kranzler (1992a) sought to replicate and extend Larson and Alderton's (1990) worst performance analysis. Whereas Larson and Alderton examined one RT task and three psychometric tests that varied in *g* saturation, Kranzler examined one highly *g*-loaded psychometric test [Multidimensional Aptitude Battery (MAB)] and three RT tasks that varied in complexity and *g* saturation. (The *g* loading of an RT task is thought to reflect complexity of information processing: more complex RT tasks have higher *g* loadings than less complex RT tasks.) Kranzler reasoned that if the WPR holds only for *g*-loaded measures, then evidence for the WPR should be obtained for highly complex RT tasks but perhaps not for less complex RT tasks.

To test his hypothesis, Kranzler (1992a) administered three RT tasks to University of California, Berkeley, students ($N = 97$; mean age = 20.3 years; mean MAB Full-Scale IQ = 120.17, S.D. = 10.88). The RT apparatus (a box with buttons) contained a home button and one or eight response buttons, which were occasionally illuminated. Participants began each trial with their hand on the home button, waited until one or more response buttons lit up, and then pressed the appropriate response button as quickly as possible. In simple RT (SRT), participants saw only one response button, waited until that button lit up, and then pressed the lit button. In choice RT (CRT), participants saw all eight response buttons, waited until one of the eight response buttons lit up, and then pressed the lit button. In Odd-Man-Out (OMO), participants saw eight response buttons, waited until three response buttons lit up simultaneously, and then pressed the "odd man" (the button that was farthest away from the other two). The three RT tasks varied in complexity. SRT was the least complex because it required no discrimination among response alternatives. CRT and OMO were moderately complex because they required discriminating among several response alternatives.

Participants' RT data were subjected to a worst performance analysis. First, each participant's RTs were ranked from fastest to slowest, separately for each RT task. Second, the RTs for each task were divided into four to seven RT bands with five RTs per band. Finally, Spearman rank-order correlations were computed between the median of each RT band and the *g* factor, which was extracted from the MAB. Results indicated that the correlations between RT bands and *g* varied with task complexity. For SRT (the least complex task), the correlations with *g* showed no linear increase across RT bands ($r_s = -.15$ and $-.13$, for the fastest and slowest RT bands, respectively). For CRT and OMO (both more complex than SRT), however, the correlations with *g* increased linearly from the fastest to the slowest RT bands ($r_s = -.16$ and $-.36$ for CRT; $-.21$ and $-.36$ for OMO). On the basis of these findings, Kranzler (1992a) concluded that the WPR varies with the complexity of an RT task, with the WPR being observed only on relatively complex (and presumably *g*-loaded) RT tasks.

2.4. *Diascro and Brody (1993)*

Diascro and Brody (1993) tested the WPR for two types of cognitive processing: serial processing, which involves processing multiple operations sequentially, and parallel processing, which involves processing multiple operations simultaneously. Diascro and Brody reasoned that if the WPR was functionally invariant, then the WPR should be found for both types of processing.

Diascro and Brody (1993) administered 384 trials of two visual-search tasks to introductory psychology students ($N = 47$; mean Cattell Culture-Fair IQ = 124.7, S.D. = 12.41). In a serial-processing task, participants had to detect the presence or absence of a vertical line surrounded by tilted lines. In a parallel-processing task, participants had to detect the presence or absence of a titled line surrounded by vertical lines. The number of distractors (1, 5, or 10) was varied across trial blocks. Previous research indicated that RTs increased monotonically with number of distractors for the serial-processing task but not for the parallel-processing task, supporting the serial versus parallel processing distinction.

Participants' RTs were subjected to a worst performance analysis. First, each participant's RTs were sorted from fastest to slowest into five percentile bands (5th, 25th, 50th, 75th, and 95th percentiles). The sorting was done separately by task (serial or parallel), target presence (present or absent), and number of distractors (1, 5, or 10 distractors), resulting in 12 different conditions with 32 trials per condition. Next, the median RT at each percentile band was correlated with IQ. Results indicated that the magnitude of the correlations in all 12 conditions became increasingly negative from the fastest to the slowest percentile bands (mean *r*s collapsing across conditions=.06, .01, $-.05$, $-.15$, and $-.21$, for 5th to 95th percentile bands, respectively). On the basis of these findings, Diascro and Brody (1993) concluded that the WPR holds for both serial- and parallel-processing tasks.

2.5. Coyle (2001)

Coyle (2001) tested whether the WPR could be generalized to a new task and a new age group. Whereas previous studies of the WPR administered RT tasks to adults, Coyle administered a word-recall task to children. Coyle reasoned that if the WPR was a general rule of cognition, then it should be observed on tasks other than RT tasks and for participants other than adults. He also argued that if the WPR applied to measures of *performance*, then it should not be observed for measures of *strategies* (i.e., preparatory behaviors preceding performance).

To test these hypotheses, Coyle (2001) gave five sort-recall trials of categorizable words to children ($N=81$; mean age = 9.4 years; mean IQ = 108.94, S.D. = 11.81). The dependent measures were recall, the number of words remembered on each trial; and strategy use, the number of memory strategies used on each trial (e.g., rehearsal or recalling words by categories). IQ was derived from the Vocabulary and Information subtests of the Wechsler Intelligence Scale for Children, Third Edition (WISC-III). The Vocabulary and Information subtests of the WISC-III have the highest *g* loadings of the WISC subtests (Jensen, 1998, p. 590).

The recall and strategy data were subjected to a worst performance analysis. First, the worst recall trial (fewest words recalled) and the best recall trial (most words recalled) were identified. Second, the fewest strategies used across trials and the most strategies used across trials were identified. Finally, Pearson correlations were computed between IQ and the measures of recall and strategy use. Results indicated that the worst recall trial predicted IQ (r=.38) better than the best recall trial (r=.13), better than the fewest strategies used across trials (r=.24), and better than the most strategies used across trials (r=.12). Furthermore, when the recall and strategy variables were regressed on IQ using the stepwise procedure, the worst recall trial was the first and only variable to be entered and it accounted for more unique variance in IQ than any other variable. On the basis of these findings, Coyle (2001) argued that the WPR was not limited to RT tasks or adult participants but appeared to be a general rule of cognition and intelligence.

2.6. Coyle (2003)

Coyle (2003) examined the relationship between the WPR and Spearman's "law of diminishing returns" (hereafter Spearman's law) (Jensen, 1998, pp. 585–588; Spearman, 1927, pp. 217–221). Spearman's law states that the *g* loading of IQ declines as participants' mental ability increases. Coyle reasoned that if the WPR is related to *g*-loaded measures, and if the *g* loading of IQ declines with mental ability, then the WPR may not hold for high ability participants, whose IQs should be low in *g* saturation.

To test this hypothesis, Coyle (2003) performed a worst performance analysis on gifted children's ($N=85$, mean WISC IQ = 140.19) word-recall data and compared the results of this analysis to those obtained in Coyle's (2001) study, which involved nongifted children. The gifted children received the same five trial word-recall task as did Coyle's (2001) nongifted participants. Coyle (2003) reasoned that if the WPR holds only for *g*-loaded measures, and if IQ is less *g* loaded for gifted children, then the WPR may not hold for gifted children.

Coyle's (2003) hypothesis was confirmed. Analyses of the gifted data indicated that the correlation between worst performance (lowest recall trial) and IQ (r=.06) did not differ significantly from zero or from the correlation between best performance (highest recall trial) and IQ (r=.15). Additional analyses indicated that the IQ-worst performance correlation for gifted children was significantly lower than the corresponding correlation for nongifted children (r=.38). Analyses of potential confounds revealed that variance compression or range restriction could not account for the results. On the basis of these findings, Coyle concluded that the WPR holds for participants whose IQs are high in g saturation (nongifted children) but not for participants whose IQs are low in g saturation (gifted children).

2.7. Salthouse (1998)

Salthouse (1998) sought to replicate and extend the WPR using different RT tasks and different cognitive ability tests. Participants aged 18 to 88 years (N = 383, mean age = 47 years) received two RT tasks. In digit–digit RT, participants saw a pair of digits on each trial and had to decide whether the digits were physically identical. In digit–symbol RT, participants saw a digit and a symbol on each trial and had to decide whether the symbol corresponded to the digit by referring to a code table. Each RT task was given on 180 trials. Participants also received cognitive tests of vocabulary knowledge, matrix reasoning, free recall, perceptual speed, and block assembly.

Participants' RTs were subjected to a worst performance analysis. First, each participant's RTs were sorted from fastest to slowest into nine percentile bands (10th, 20th, 30th, 40th, 50th, 60th, 70th, 80th, and 90th percentiles). Next, the mean RT for each percentile band was computed. Finally, correlations between the percentile bands and the cognitive ability tests were computed, controlling for age. Contrary to the WPR, results indicated that the correlations declined slightly (in most cases) from the 10th to the 30th percentiles and stabilized (or increased slightly) thereafter (Salthouse, 1998, Figs. 3 and 6, pp. 161 and 164).

Salthouse's (1998) findings cannot easily be interpreted because he combined data from different age groups and analyzed cognitive test data over a large age range (18 to 88 years). This is a problem because the g loadings of cognitive tests may decline with age from early to later adulthood (Deary et al., 1996; Jensen, 1998, pp. 585–588; Spearman, 1927, pp. 217–221). In Salthouse's study, a decline in the g loadings of the cognitive tests would have reduced the likelihood of obtaining evidence for the WPR, which applies only to g-loaded measures. One solution would be to compute RT–IQ correlations separately at each age level so that the g loadings of the cognitive tests are independent for each age group. Until such analyses are performed, Salthouse's findings cannot be interpreted as providing evidence against the WPR.

2.8. Summary of evidence

With the exception of Salthouse's (1998) study, support for the WPR was widespread and robust. Specifically, support for the WPR was obtained for children as well as adults; for RT measures as well as memory measures; for serial-processing tasks as well as parallel-processing tasks; and for visual stimuli as well as auditory stimuli. It is worth noting that several of the studies reviewed reported multiple and independent analyses of the WPR (e.g., Coyle, 2001; Diascro & Brody, 1993; Jensen, 1982; Kranzler, 1992a; Larson & Alderton, 1990). Although the results of any analysis could have rejected the WPR, the results consistently supported the WPR. It is also worth noting that evidence for the WPR was obtained despite considerable variation in tasks, participants, and measures, minimizing the likelihood that the WPR was an artifact of these factors.

3. Broader implications of the WPR

The WPR warrants attention for at least three reasons. First, the WPR appears to contradict certain assumptions of classical measurement theory. Second, the WPR and a worst performance analysis raise new questions about intelligence that traditional methods cannot address. Finally, the WPR appears to vary with psychometric *g*. Each of these issues will be discussed below.

3.1. The WPR and measurement theory

Classical measurement theory (CMT) states that all scores contain two types of variance: true-score variance, which is reliable, and error variance, which is unreliable (Crocker & Algina, 1986). CMT further assumes that when multiple scores are obtained for a participant, scores close to the central tendency of the participant's distribution should be loaded with true-score variance and scores deviating sharply from central tendency should be loaded with error variance. It follows from these assumptions that extreme scores in a participant's distribution should contain more error variance than the average score in a participant's distribution.

The assumptions of CMT have implications for the WPR. If, as CMT maintains, extreme scores in a participant's distribution contain more error than the average score in a participant's distribution, then worst performance (an extreme score) should predict IQ more poorly than average performance (for a similar argument, see Baumeister, 1998). Research on the WPR contradicts this prediction. Larson and Alderton (1990), for example, found that the correlation between the slowest RT band and IQ ($r_s = -.37$) was slightly higher in magnitude than the correlation between the median RT and IQ ($r_s = -.36$). Similarly, Kranzler (1992a) found that the correlation between the slowest RT band and IQ was higher in magnitude than the correlation between the median RT and IQ for CRT ($r_s = -.36$ vs. $-.20$) and for OMO ($r_s = -.36$ vs. $-.24$) but not for SRT ($r_s = -.13$ vs. $-.15$), for which the WPR was not predicted. Finally, Coyle (2001) found that the correlation between the lowest recall trial and IQ (r=.38) was higher than the correlation between mean recall and IQ (r=.30).

These findings contradict the assumption that worst performance is loaded with error. If that assumption were true, then worst performance should have been a relatively poor predictor of IQ. But the evidence indicated that worst performance consistently predicted IQ as well as, or better than, average performance. The evidence therefore suggests that worst performance is loaded with one or more types of true-score variance. The source or sources of this true-score variance have not been identified but may be related to cognitive factors (attention or working memory lapses) or to biological factors (neural errors and oscillations). Further elaboration of these possibilities will be discussed in the Section 4, ‘‘Theories of the WPR.’’

3.2. The WPR and methods for studying mental ability on multitrial tasks

The WPR and a worst performance analysis encourage researchers to apply new analytical approaches to old data and paradigms. Coyle’s (2001) worst performance analysis of children’s word recall is a good example. Coyle’s initial motivation was not to examine the WPR but to study trial-to-trial changes in recall and strategy use (see Coyle & Bjorklund, 1997). Coyle later realized that his data might be used to evaluate the WPR, performed a worst performance analysis on the archival data, and found evidence for the WPR. Kranzler’s (1992a) study of adults’ RTs is another example. Kranzler’s study involved an RT apparatus and a series of RT paradigms that had been used widely in prior studies (for a review of these studies, see Jensen, 1982). Kranzler realized that these RT paradigms might be used to test the WPR, performed a worst performance analysis using the RT paradigms, and found evidence for the WPR.

Conventional approaches for studying trial-to-trial deviations in performance cannot be used to test the WPR. Such approaches typically give participants a multitrial cognitive task, measure their performance on each trial, and compute a measure of trial-to-trial deviations from central tendency. This approach is inadequate for evaluating the WPR because it involves the computation of an aggregate measure of trial-to-trial deviations, typically the standard deviation in performance over trials. The standard deviation in performance over trials reflects variability in performance attributed to worst performance, best performance, and all performance scores in between. Thus, the standard deviation in performance over trials provides an overall estimate of trial-to-trial variability but provides no information about the unique variance attributed to worst performance. To evaluate the WPR, the standard deviation in performance over trials must be decomposed into its constituent parts (e.g., worst performance trials and best performance trials), and its constituent parts must then be subjected to a worst performance analysis.

3.3. The WPR and psychometric g

Psychometric *g* refers to the variance shared by a diverse set of mental ability tests. The *g* loading of any mental ability test can be determined by factor analysis (for a review of factor analytic studies of *g*, see Jensen, 1998).

Not all mental ability tests have equivalent *g* loadings (Jensen, 1998). Mental ability tests that have relatively high *g* loadings tend to involve mental manipulation of symbols or deduction of laws, rules, or principles. Mental ability tests that have relatively low *g* loadings tend to involve rote memorization or perceptual speed. Forward- and backward-digit span illustrate the distinction between less and more *g*-loaded mental tests. Forward-digit span (a less *g*-loaded mental test) involves recalling a list of digits in the order in which the digits are presented. Backward-digit span (a more *g*-loaded mental test) involves recalling a list of digits in the reverse order in which the digits are presented. Backward-digit span has a higher *g* loading because it requires the mental manipulation of digits from forward to reverse order prior to recall.

The WPR has been found to hold for measures high in *g* saturation but not for measures low in *g* saturation. Larson and Alderton (1990), for example, found that the WPR held for highly *g*-loaded psychometric measures (the ASVAB and a working memory test) but not for a less *g*-loaded psychometric measure (clerical speed). Similarly, Kranzler (1992a) found that the WPR held for two highly complex and *g*-loaded RT tasks (CRT and OMO) but not for the less complex SRT task. Likewise, Coyle (2001) found that the WPR held for the number of words recalled but not for the number memory of strategies used, a finding consistent with the assumption that strategies are not very *g* loaded (Alderton & Larson, 1994). Finally, Coyle (2003) found that the WPR held for nongifted children, who presumably had highly *g*-loaded IQs, but not for gifted children, who presumably had less *g*-loaded IQs.

That the WPR holds only for highly *g*-loaded measures has an interesting implication. If worst performance on a multitrial task is related to *g*-loaded measures more strongly than best performance, then worst performance should have a higher *g* loading than best performance. If this is so, then the best estimate of *g* on any multitrial cognitive task should be worst performance. Further, since the *g* loading of a mental ability test is the "active ingredient" of the test (i.e., that which gives the test its predictive utility), it follows that worst performance should be a better predictor of practical outcomes such as academic or job performance than should best performance. A study correlating worst and best performances with IQ as well as practical outcomes such as academic or job performance is needed to test this hypothesis.

4. Theories of the WPR

What causes the WPR? Larson and Alderton (1990) focused on cognitive causes of the WPR. Coyle (2001) focused on cognitive causes, too, but added that these causes have a neurological basis. Jensen (1992) developed a theory of RT linking neural physiology and cognition that has implications for the WPR. These theories are reviewed below.

In the penultimate paragraph of their article, Larson and Alderton (1990) focused on a cognitive explanation of the WPR. They derived their theory from Larson and Saccuzzo's (1989) theory of working memory. First, they argued that people occasionally experience lapses in working memory, which may interfere with chaining together representations in working memory. Second, they argued that on RT tasks these lapses in working memory produce especially slow RTs. Finally, they argued that lapses in working memory occur more frequently for low-IQ individuals than for high-IQ individuals. If, as Larson and Alderton maintain, low-IQ individuals experience more lapses in working memory than high-IQ individuals, and if these lapses produce especially slow RTs, then low-IQ individuals should show a lot of slow RTs. This prediction is consistent with evidence that the worst RT trials reveal more about intelligence than do other portions of the RT distribution.

Related to the working memory hypothesis is an attention lapse hypothesis (cf. Jensen, 1992), which proposes that attention lapses prevent task relevant stimuli from entering working memory. An important assumption of this hypothesis is that the effect of attention lapses on RT depends on when they occur. If attention lapses occur during the presentation of the reaction stimulus (RS), then the RS should not immediately be processed and RTs should be slowed. In contrast, if attention lapses occur prior to the presentation of the RS, then the RS should quickly be processed and RTs should be relatively fast. On the basis of prior research, Jensen (1992) suggested that attention lapses should be more common for low-IQ individuals than for high-IQ individuals. If Jensen's suggestion is correct, then low-IQ individuals should frequently show especially slow RTs. This prediction would be consistent with the predictions of the WPR.

Coyle (2001) argued that the WPR has a neural basis. Coyle's argument involved three claims. First, he argued that all individuals experience occasional errors of neural transmission, which produce inefficient communication among neurons. Second, he claimed that these "neural errors" occur more frequently in low-IQ individuals, who lack the necessary structural features (e.g., myelination) to reduce such errors (for a similar argument, see Miller, 1994). Finally, he argued that the neural errors have cognitive consequences, which include pronounced cognitive deficits and RT slowing. Coyle added that since low-IQ individuals experience a disproportionate number of neural errors, they should show more pronounced cognitive deficits, which may be expressed as very slow RTs.

Jensen (1982, 1992) proposed a theory of neural physiology that has implications for the WPR. Jensen noted that neuronal firing involves three sequential stages: a subthreshold period, neuronal firing (i.e., the generation of an action potential), and a refractory period. In the subthreshold period, the neuron generates excitatory potentials, which depolarize the neuron. In the neuronal firing period, the neuron depolarizes to a specified level (called threshold) and an action potential is triggered. In the refractory period, the neuron cannot generate another action potential and, in theory, cannot respond (or respond efficiently) to external stimuli. This cycle of excitation, firing, and refraction may repeat indefinitely. A consequence of the cycle is that neurons oscillate between firing (when an action potential is generated) and not firing (when the subthreshold or refractory period occurs).

Jensen (1992) argued that neural oscillations through the cycle of excitation, firing, and refraction might be linked to individual differences in RT and intelligence. His argument involved three claims. First, he noted that if the RS in an RT task occurs during the subthreshold or refractory period, the neuron would not fire. The failure of a neuron to fire immediately in response to the RS should delay responding and produce slow RTs. Second, he argued that there are individual differences in the speed of neural oscillations. Specifically, he argued that some individuals cycle through the subthreshold and refractory periods faster than others and that such individuals have a wider window in which the RS can activate neurons. These individuals should be able to respond to the RS more quickly and produce faster RTs. Finally, Jensen argued that there is a link between individual differences in neural oscillations and IQ such that individuals whose neurons cycle through the subthreshold and refractory periods more quickly have relatively high IQs.

Jensen's (1992) theory has implications for the WPR. If low-IQ individuals spend more time in the subthreshold and refractory periods, and if neural (and presumably behavioral) responses are not possible during these periods, then low-IQ individuals should be less likely to respond immediately to an RS. This may explain why low-IQ individuals show especially slow RTs. If, however, the RS happens to coincide with the period when neurons can fire, then low-IQ individuals should show fast neural and behavioral responses. This may explain why low-IQ individuals periodically show really fast RTs even though most of their RTs are very slow. High-IQ individuals, in contrast, should need less time to cycle through the subthreshold and refractory periods, and so their neurons should be more likely to respond immediately to the RS and their RTs should be relatively fast. This may explain why high-IQ individuals show less of a difference between their fastest and slowest RTs (Baumeister & Kellas, 1968).

5. Alternative hypotheses for the WPR

Researchers have raised the possibility that the WPR may be attributed to statistical or data artifacts such as measurement unreliability, variance compression, or outliers. In the pages that follow, the possibility that the WPR can be attributed to statistical or data artifacts is examined. Specifically, a default hypothesis linking the WPR and *g* will be presented and a number of alternative hypotheses will be examined. A review of the alternative hypotheses will show that they cannot account for the data as well as the default hypothesis.

5.1. The default hypothesis

Several studies (Coyle, 2001, 2003; Jensen, 1982; Kranzler, 1992a; Larson & Alderton, 1990) have examined the WPR for cognitive tasks with different *g* loadings. The findings of these studies indicate that evidence for the WPR increases with the *g* loading of a task. These findings represent the default hypothesis for the WPR, which proposes that the WPR is more likely to hold for cognitive tasks high in *g* saturation than for cognitive tasks low in *g* saturation.

Evidence for the default hypothesis is quite robust. Larson and Alderton (1990), for example, found that the WPR was indicated for a measure of working memory, a highly *g*-loaded construct, but not for a measure of clerical speed, a less *g*-loaded construct. Similarly, Kranzler (1992a) found that the WPR was indicated for CRT and OMO, two complex RT tasks with high *g* loadings, but not for SRT, a less complex RT task with a low *g* loading. Likewise, Coyle (2001) found that the WPR was indicated for a measure of word recall, a task relying heavily on working memory (a highly *g*-loaded construct), but not for memory strategy use, a measure believed to be low in *g* saturation. Finally, Coyle (2003) also found that the WPR was indicated for nongifted children, who have highly g-loaded IQs, but not for gifted children, who have less g-loaded IQs.

The results of these studies converge on a simple rule: the WPR is more likely to hold for measures high in g saturation than for measures low in g saturation. Researchers have examined other explanations for the WPR that are related to data or statistical artifacts. These explanations will be reviewed below as a series of alternative hypotheses.

Hypothesis 1: Can outliers explain the WPR?

One alternative hypothesis for the WPR concerns outliers, extreme scores lying outside the normal range of values. One or a few outliers could produce evidence for the WPR. For example, a worst performance outlier could inflate the magnitude of the correlation with g, or a best performance outlier could decrease the magnitude of the correlation with g. Either type of outlier would increase the likelihood of obtaining evidence for the WPR.

One argument against the alleged outlier problem is that evidence for the WPR is still obtained when within-participant outliers are removed. Coyle (2001), for example, omitted each participant's highest and lowest recall scores, selected the next highest and lowest recall scores for each participant, and then correlated these scores with a measure of g. Coyle found that worst performance still predicted g better than best performance even after the outliers had been removed. Larson and Alderton (1990) performed an analogous analysis with RT data. They omitted each participant's fastest and slowest RTs, created 16 RT bands (of five RTs each), and then correlated each RT band with a measure of g. They found that the magnitude of the correlations with g increased monotonically from the fastest to the slowest RT bands even after the outliers had been omitted.

Another argument against the possibility that the WPR exploits outliers comes from Larson and Alderton's (1990) study. As mentioned, Larson and Alderton computed correlations between each of 16 RT bands and g and found that the correlations increased linearly from the fastest to the slowest RT bands. On the basis of these results, Larson and Alderton argued that the RT band outliers (i.e., the fastest and slowest RT bands) could be omitted and the WPR would still be supported. In Larson and Alderton's words, "The data are lawful; outliers are not" (p. 319).

Hypothesis 2: Can variance compression explain the WPR?

Another alternative hypothesis for the WPR concerns variance compression. If variance compression were indicated for best performance but not for worst performance, then correlations with best performance would necessarily be lower than correlations with worst performance (assuming all other variables were held constant), which is exactly what the WPR predicts.

Data that appear to be consistent with the variance compression argument come from Larson and Alderton (1990). Larson and Alderton found that as the correlations with g increased from the fastest to the slowest RT bands, so, too, did the standard deviations for the RT bands. Larson and Alderton acknowledged that the additional RT variability for the worst performance bands could increase the magnitude of the correlations with worst performance. They countered, however, that these differences in intersubject variability were not a confound but an answer. They argued, "If the question is, 'On what aspect of the data should we focus our study of individual differences?', the answer is, 'Obviously, to the place where those differences emerge'" (p. 319).

There are additional arguments against the alleged variance compression problem. One of these arguments is that evidence for the WPR can be obtained when using measures of association that are relatively insensitive to differences in intersubject variability. For example, evidence for the WPR has been obtained when using Spearman's rank-order correlation (Kranzler, 1992a; Larson & Alderton, 1990), which, compared to Pearson's correlation, is less dependent on differences in intersubject variability. Such evidence minimizes the potential confound of variance compression in studies of the WPR.

Another argument against the alleged variance compression problem is that evidence for the WPR is obtained when no variance compression is observed. Kranzler (1992a), for example, found that for CRT, correlations between RT and *g* increased monotonically from the fastest to the slowest RT bands, even though the standard deviations did not increase consistently across RT bands (S.D.s =.41, .53, .53, .57, .53, and .55, for the fastest to slowest RT bands, respectively). Similarly, Coyle (2001) found that on a word-recall task worst performance (lowest recall trial) predicted a *g*-loaded measure better than best performance (highest recall trial), even though the standard deviations for worst and best performances were nearly identical (S.D. = 3.09 and 2.86, for minimum and maximum recall, respectively). The fact that evidence for the WPR is obtained even when variability for worst and best performances is equivalent (or nearly so) greatly minimizes the alleged variance compression problem.

Hypothesis 3: Can skewness explain the WPR?

Another alternative hypothesis for the WPR concerns the skewness of performance distributions. Most research on the WPR examines RT. RT distributions are almost always positively skewed (skewed to the right), with many fast RTs and few slow RTs. One reason for the asymmetric shape of the RT distribution has to do with constraints set by human physiology (Jensen, 1992). RTs generally cannot be lower than 100 ms, the minimum amount of time needed to recognize and respond to an external stimulus. In contrast, RTs have no physiological upper bound. The result is that RTs tend to bunch around the fast end of the distribution and spread out around the slow end of the distribution.

The positive skew of the RT distribution might be related to the WPR. In particular, the wide spread of scores at the slow end of the RT distribution might provide greater sensitivity for detecting individual differences in IQ. Conversely, the compressed spread of scores at the fast end of the RT distribution might provide less sensitivity. Both conditions could enhance the likelihood of obtaining evidence for the WPR.

Two studies have examined the relation between RT skewness and IQ. Juhel (1993) found that RT skewness on a visual discrimination task was significantly related to IQ ($r = -.40$, $N = 52$). In contrast, Kranzler (1992b) found that RT skewness on three RT tasks was *not* significantly related to IQ (rs=.15, $-.06$, $-.16$, for SRT, CRT, and OMO RT, respectively;

$N = 101$). These discrepant findings may be related to the number of RT trials used in each study: Juhel used 128 RT trials whereas Kranzler did not report the number of RT trials used. It is possible that Kranzler used too few RT trials to detect significant correlations between RT skewness and IQ.

Although neither Juhel (1993) nor Kranzler (1992b) directly examined the relation between the WPR and RT skewness, it would not be surprising to find a link between the WPR and RT skewness. Research on the WPR has demonstrated that low-IQ participants frequently show especially slow RTs (Baumeister & Kellas, 1968; Jensen, 1982; Larson & Alderton, 1990). This finding suggests that low-IQ participants should have RT distributions with especially long tails to the right of the distribution, indicating exaggerated positive skew. High-IQ participants, in contrast, rarely show slow RTs and so their RT distributions should be less positively skewed.

Because RT distributions are almost always positively skewed, it would be useful to demonstrate the WPR for a performance distribution that was relatively symmetrical (and not skewed). A reanalysis of Coyle's (2001) word-recall data provides such a demonstration. Recall that Coyle analyzed a measure of performance (number of words recalled) that, unlike RT, is usually normally distributed. Recall also that Coyle found that worst performance predicted IQ better than best performance, which is evidence for the WPR.

Coyle's (2001) data were reanalyzed here for degree of skewness. Skewness statistics were computed for measures of worst performance (lowest recall trial), best performance (highest recall trial), and IQ. A skewness statistic of 0 indicates the distribution is perfectly symmetrical. Positive skew values indicate the distribution is skewed to the right; negative skew values indicate the distribution is skewed to the left. A general rule of thumb is that a distribution is relatively symmetrical (and not skewed) if its skewness values are between -1 and 1. The skewness values for the variables in Coyle's study were between -1 to 1 (skewness=.45, $-.07$, and $-.19$, for worst performance, best performance, and IQ, respectively).

An additional reanalysis of Coyle's (2001) data computed skewness values for each participant's distribution of recall scores across the five trials. This approach is perhaps more appropriate because in a worst performance analysis, worst and best performance scores are selected from each participant's distribution, not from the group's distribution. The reanalysis indicated a relatively normal distribution of recall scores across trials (mean skewness $= -.19$). Further, the skewness of each participant's recall scores was not significantly correlated with IQ ($r = -.16$, $N = 81$, $P > .10$). These data argue against the skewness hypothesis. If that hypothesis had merit, then evidence for the WPR should have been found for skewed distributions only. But this was not the case. The data presented here show that evidence for the WPR was obtained for relatively symmetrical distributions. These findings suggest that the WPR holds for both normal and skewed distributions, suggesting that the shape of the performance distribution is not necessarily related to the WPR.

Hypothesis 4: Can measurement reliability explain the WPR?

Another alternative hypothesis for the WPR concerns measurement reliability (or lack thereof). If best performance scores are less reliable than worst performance scores, then correlations with best performance should be lower in magnitude than correlations with worst performance (all other things being equal). Such a situation would produce evidence for the WPR as a result of differences in measurement reliability.

There are two ways to evaluate the alleged reliability problem. The first is to demonstrate the reliability of worst and best performances using conventional techniques (e.g., test–retest reliability). Unfortunately, such techniques have not been reported in the extant literature. A second approach is to compute validity coefficients for worst and best performances and infer the reliability of each performance measure. If best performance really is less reliable than worst performance, then correlations with best performance should be lower than correlations with worst performance for all outcomes, including (but not limited to) *g*-loaded measures.

Larson and Alderton (1990) provide evidence against the alleged unreliability of best performance. They computed correlations between each participant's 16 RT bands and median RT over trials. They found that the correlation of median RT with the fastest RT band (r_s=.80) was nearly identical in magnitude to the correlation of median RT with the slowest RT band (r_s=.82). More generally, the correlations of median RT with the faster versus slower RT bands were of comparable magnitude throughout the range of RT bands. For example, the correlation with the second fastest RT band (r_s=.88) was similar in magnitude to the correlation with the second slowest RT band (r_s=.90); the correlation with the third fastest RT band (r_s=.92) was similar in magnitude to the correlation with the third slowest RT band (r_s=.93); and so on. These data argue against the alleged unreliability of best performance. If that argument had merit, then correlations with best performance should have been lower (in magnitude) than correlations with worst performance. But this was not so: The correlations with the fastest and slowest RTs were comparable in magnitude throughout the range of RT bands.

Kranzler (1992a) provides a different kind of evidence against the alleged unreliability of best performance. You may recall that Kranzler performed a worst performance analysis for three RT tasks: SRT, CRT, and OMO. As predicted, Kranzler found that the more complex RT tasks (CRT and OMO) revealed evidence for the WPR whereas the least complex RT task (SRT) revealed no such evidence. The SRT results indicated that the correlation with best performance ($r_s = -.15$) was comparable in magnitude to the correlation with worst performance ($r_s = -.13$). These data seem more consistent with Kranzler's hypothesis that the WPR varies with an RT task's *g* loading than the hypothesis that best performance is unreliable. If best performance was unreliable, then correlations with best performance should have been lower (in magnitude) than correlations with worst performance for all RT tasks. But this was not so: The magnitude of correlations with best and worst performances varied as a function of each RT task's putative *g* loading, with greater differentiation between best and worst performance correlations on tasks believed to be high in complexity (and presumably *g*).

Coyle (2001) provides yet another kind of evidence against the argument that best performance is unreliable. You may recall that Coyle computed correlations between best performance (highest recall trial), worst performance (lowest recall trial), and IQ. As predicted, Coyle found that the results supported the WPR (i.e., the lowest recall trial predicted IQ better than the highest recall trial). Coyle also computed correlations between best performance, worst performance, and strategy use (mean number of memory strategies used across trials). With respect to these latter analyses, Coyle found that the correlation between best performance and strategy use (r=.66) was nearly identical in magnitude to the correlation between worst performance and strategy use (r=.65). These data argue against the hypothesis that best performance is unreliable. If that hypothesis were true, then correlations with best performance should have been lower than correlations with worst performance for all outcome measures. But this was not always the case: The correlations with best performance were comparable to those with worst performance when strategy use (but not IQ) was being predicted.

Hypothesis 5: Does the WPR vary with trial?

Another alternative hypothesis for the WPR concerns a confound between worst performance and trial novelty. Sternberg (1985) argued that responses on relatively novel trials reveal more about intelligence than do responses on relatively familiar trials. If responses on novel trials are a good index of intelligence, and if worst performance is typically observed on the first few trials (i.e., the most novel trials), then the WPR may be confounded with novelty of trial. Such a confound would make it difficult to disentangle the effect of worst performance and the effect of novelty of trial.

Extant research on the WPR does not provide sufficient data to evaluate the alleged confound. For example, neither Larson and Alderton (1990) nor Kranzler (1992a) reported whether the slowest RTs were randomly distributed across trials or were concentrated on the early (and most novel) trials. Similarly, Coyle (2001) did not report whether the lowest recall scores were observed primarily on the early trials or on the later trials. Without knowing when the worst performance trials were observed, it is impossible to disentangle the effects of worst performance and the effects of trial novelty.

In an effort to evaluate the alleged confound between worst performance and trial novelty, a reanalysis of Coyle's (2001) data was performed here. First, the number of times worst performance (the lowest recall score) was observed on each of the trials was tabulated. (Only data from participants who had a unique minimum recall score were used.) Second, the resulting frequencies were analyzed to determine if the distribution of worst performance scores varied across trials. The analysis revealed that the distribution of worst performance scores was comparable across trials, Cochran's $Q = 7.094$, $N = 64$, $df = 4$, $P > .10$ (percentage of worst performance scores = 17, 13, 16, 23, and 31, for Trials 1 to 5, respectively). These data argue against the alleged confound between worst performance and trial novelty. If that confound had merit, then the worst performance scores should have been concentrated on the early trials. But this was not the case: The worst performance scores were distributed fairly evenly across trials, with a slight (but nonsignificant) tendency for the worst performance scores to be found on the later trials.

In sum, a review of alternative hypotheses indicates that the WPR cannot readily be attributed to statistical or data artifacts such as outliers, variance compression, or measurement unreliability. Instead, the preponderance of evidence indicates that the WPR varies as a function of a task's *g* loading, which is evidence for the default hypothesis.

6. Future directions for WPR research

This section explores four possibilities for future investigations of the WPR: (a) the physiological correlates of the WPR, (b) the WPR and nonnormative subject populations, (c) the heritability of worst and best performances, and (d) the correlates of best performance.

6.1. Physiological correlates

One theory of the WPR is that individual differences in IQ are associated with errors in brain processes, which express themselves as lapses in attention or working memory. Jensen (1982, 1992) has suggested that these errors in brain processes are manifested as different patterns of brain activity and are observed primarily on worst performance trials. If Jensen's suggestion is correct, then participants with different IQ levels should show different patterns of brain activity on worst performance trials but perhaps not on best performance trials. One way to test this hypothesis would be to measure event-related potentials (ERPs) of the brain while high- and low-IQ participants perform an RT task over trials. The ERP and RT data could then be subjected to a 2 (participant IQ: high vs. low) $\times$ 2 (RT performance: worst vs. best) worst performance analysis. On the basis of extant theory of brain–behavior relationships (Gevins, 2002), levels of brain activity in areas related to working memory and attention (e.g., prefrontal cortex) should discriminate between high- and low-IQ participants on worst performance trials but perhaps not on best performance trials.

6.2. Nonnormative subject populations

Attention deficit hyperactivity disorder (ADHD) provides an opportunity to test the attention lapse hypothesis of the WPR. ADHD participants show a variety of cognitive deficits including frequent and sustained attention lapses. The effect of these lapses on ADHD participants' RTs may vary with the presentation of an RS. For example, if the lapses occur during the presentation of the RS, then ADHD participants should not immediately process the RS and should show delayed RTs. In contrast, if the lapses occur prior to (but not during) the presentation of the RS, then ADHD participants should quickly process the RS and should show faster RTs. A test of these hypotheses might involve administering an RT task repeatedly to ADHD and normal participants. The key prediction would be that, relative to normal participants, ADHD participants should frequently show slow RTs (due to their frequent attention lapses) but occasionally show fast RTs (due to the occasional nature of their lapses). Furthermore, if the underlying cause of the WPR is attention lapses and not low IQ, then the predicted results should be obtained when ADHD and normal participants are equated on IQ.

6.3. Heritability

There is considerable evidence that the heritability coefficient (h^2) for a cognitive task increases with a task's *g* loading (Jensen, 1998, p. 184). The relation between a task's *g* loading and its heritability coefficient has implications for the WPR. If, as the WPR maintains, worst performance has a higher *g* loading than best performance, then the heritability coefficient for worst performance should be higher than the heritability coefficient for best performance. Further, if worst performance functions as other *g*-loaded measures whose heritability coefficients increase over the lifespan, then the magnitude of the heritability coefficient for worst performance should increase as well.

6.4. Predictors of best performance

Although a worst performance analysis focuses on the correlates of worst performance, the same sort of analysis can yield important insights about the correlates of *best* performance. A reanalysis of Coyle's (2001) data illustrates this point. Recall that Coyle gave five word-recall trials to children in second through fourth grade (N = 81, mean age = 9.4 months). Recall also that Coyle correlated worst and best performances (lowest and highest recall scores) with IQ and found evidence for the WPR (i.e., worst performance was the best predictor of IQ). The reanalysis of Coyle's data performed here correlated worst and best performances with *grade level*.

It was predicted that grade level would be more strongly related to best performance (most words recalled) than to worst performance (fewest words recalled). This prediction was based on the assumption that developmental changes in memory capacity would place a ceiling on the maximum number of words recalled at each grade, making best performance a good predictor of grade level. In contrast, memory capacity should place no floor on the minimum number of words recalled: Both older and younger children occasionally show very low levels of recall because of attention lapses and processing errors and, in such cases, worst performance would not discriminate well between grade levels.

The results confirmed these hypotheses. First, best performance predicted grade (r=.44) better than worst performance (r=.35). Second, when best and worst performances were regressed on grade using the stepwise procedure, best performance was the first and only variable to be entered, $F(1,79) = 18.92$ (βs=.44 and .12, for best and worst performances, respectively). Finally, forcing worst performance into the model at the first step did not prevent best performance from being entered at the second step, $F(2,78) = 9.81$ (βs=.12 and .37, for worst and best performances, respectively). (A similar pattern of results was obtained when age was substituted for grade in the analyses.) These results suggest that best performance predicts measures of developmental level better than worst performance. Moreover, the results demonstrate that a worst performance analysis can provide insights into the correlates of best performance as well as insights into the correlates of worst performance.

7. Summary

The purpose of this paper was to review research and theory on the WPR, which states that on multitrial cognitive tasks, worst performance trials predict general intelligence better than best performance trials. The following were the main findings:

1. Support for the WPR was widespread and robust. In particular, support for the WPR was found for children as well as adults; for RT measures as well as memory measures; for serial-processing tasks as well as parallel-processing tasks; and for visual stimuli as well as auditory stimuli.
2. The WPR challenged conventional assumptions concerning the measurement of psychological and behavioral responses. In particular, the WPR challenged the assumption that extreme scores are loaded with error, and the corollary that extreme scores should not predict outcomes as well as central tendency scores.
3. The WPR had a plausible cognitive and biological basis. One theory held that the WPR might be related to lapses in attention and working memory. Another theory held that the WPR might be related to individual differences in the frequency of neural errors and oscillations.
4. The WPR could not readily be explained by statistical or data artifacts. In particular, the WPR could not be attributed to outliers, variance compression, or unreliable measurement. While any of these artifacts could have explained the WPR, none held up under scrutiny.
5. The WPR varied with the g loading of a cognitive task. In particular, the WPR held for cognitive tasks high in g saturation but not for cognitive tasks low in g saturation.

While the preponderance of evidence suggests that the WPR cannot be attributed to statistical or data artifacts, some problems still remain. A central problem concerns the causes of the WPR. Future research needs to move beyond simple description of the WPR and explore its cognitive and biological basis. Another problem concerns the causes of *best* performance. A worst performance analysis examines the full range of scores in an individual's performance distribution. It therefore raises questions about the causes of *best* performance as well as the causes of *worst* performance.

Acknowledgements

This research was supported by a Faculty Development Leave at the University of Texas at San Antonio, and by a Visiting Scientist Fellowship at the Research Imaging Center, University of Texas Health Science Center at San Antonio. I would like to thank Deborah

Coyle, Douglas Detterman, Stephen Petrill, Arthur Jensen, and an unnamed reviewer for their thoughtful comments.

References

Alderton, D. L., & Larson, G. E. (1994). Cross-task consistency in strategy use and the relationship with intelligence. *Intelligence*, *18*, 47–76.

Baumeister, A. A. (1998). Intelligence and the "personal equation". *Intelligence*, *26*(3), 255–265.

Baumeister, A. A., & Kellas, G. (1968). Distributions of reaction times of retardates and normals. *American Journal of Mental Deficiency*, *72*, 715–718.

Coyle, T. R. (2001). IQ is related to the worst performance rule in a memory task involving children. *Intelligence*, *29*, 117–129.

Coyle, T. R. (2003). IQ, the worst performance rule, and Spearman's law: A reanalysis and extension. *Intelligence*, *31*, 473–489.

Coyle, T. R., & Bjorklund, D. F. (1997). Age differences in, and consequences of, multiple- and variable-strategy use on a multitrial sort-recall task. *Developmental Psychology*, *33*, 372–380.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart, and Winston.

Deary, I. J., Egan, V., Gibson, G. J., Austin, E. J., Brand, C. R., & Kellaghan, T. (1996). Intelligence and the differentiation hypothesis. *Intelligence*, *23*, 105–132.

Diascro, M. N., & Brody, N. (1993). Serial versus parallel processing in visual search tasks and IQ. *Personality and Individual Differences*, *14*, 243–245.

Gevins, A. (2002). Electrophysiological imaging of brain function. In A. W. Toga, & J. C. Mazziotta (Eds.), *Brain mapping: The methods* (2nd ed.) (pp. 175–86). San Diego, CA: Academic Press.

Jensen, A. R. (1982). Reaction time and psychometric *g*. In H. J. Eysenck (Ed.), *A model for intelligence* (pp. 93–132). New York: Springer.

Jensen, A. R. (1992). The importance of intraindividual variability in reaction time. *Personality and Individual Differences*, *13*, 869–882.

Jensen, A. R. (1998). *The g factor: The science of mental ability*. Westport, CT: Praeger.

Juhel, J. (1993). Should we take the shape of reaction time distributions into consideration when studying the relationship between RT and psychometric intelligence? *Personality and Individual Differences*, *15*, 356–360.

Kranzler, J. H. (1992a). A test of Larson and Alderton's (1990) worst performance rule of reaction time variability. *Personality and Individual Differences*, *13*, 255–261.

Kranzler, J. H. (1992b). The skewness of the distribution of RT trials does not correlate with psychometric *g*. *Personality and Individual Differences*, *13*, 945–946.

Larson, G. E., & Alderton, D. L. (1990). Reaction time variability and intelligence: A "worst performance" analysis of individual differences. *Intelligence*, *14*, 309–325.

Larson, G. E., & Saccuzzo, D. P. (1989). Cognitive correlates of general intelligence: Toward a process theory of *g*. *Intelligence*, *13*, 5–31.

Miller, E. M. (1994). Intelligence and brain myelination: A hypothesis. *Personality and Individual Differences*, *17*, 803–832.

Salthouse, T. A. (1998). Relations of successive percentiles of reaction time distributions to cognitive variables and adult age. *Intelligence*, *26*, 153–166.

Spearman, C. (1927). *The abilities of man: Their nature and measurement*. New York: Macmillan.

Sternberg, R. J. (1985). *Beyond IQ: A triarchic theory of human intelligence*. New York: Cambridge University Press.