Running Head: The Magical Number Seven

The Magical Number Seven:
What Did Miller Really Prove?

Bruce L. Bachelder

Bachelder & Bachelder

Psychological & Educational Services

Morganton, NC 28655-3729

Correspond:

Bruce L. Bachelder

306 West Union St

Morganton, NC 28655-3729

bruce@brucebachelderphd.com

Abstract

It is widely but erroneously believed (e.g. Baddeley, 1994; Shiffrin & Nosofsky, 1994) that Miller (1956) proved the three *magical numbers* are coincidental. This erroneous interpretation constitutes a barrier to the wider evaluation of span theory, a unitary theory of the magical numbers (Bachelder & Denny, 1977a, 1977b). Miller suspected coincidence, but withheld judgment (p. 96). His paper actually asserted, evaluated, then rejected one unitary explanation of them, the *channel capacity hypothesis*. Rejection of a single explanatory hypothesis does not prove coincidence. This issue is important and the limitations need to be explained (Cowan, 1999, 2001; Shiffrin and Nosofsky, 1994, p. 360).

The Magical Number Seven: What Did Miller Really Prove?

> What about the magical number seven? What about the seven wonders of the world, the seven seas, the seven deadly sins, the seven daughters of Atlas in the Pleiades, … the seven categories for absolute judgment, the seven objects in the span of attention, and the seven digits in the span of immediate memory? For the present *I propose to withhold judgment* [italics added]. Perhaps there is something deep and profound behind all these sevens, something just calling out for us to discover it. But I *suspect* [italics added] that it is only a pernicious, Pythagorean coincidence. (Miller, 1956, p. 96)

These sentences are from the concluding paragraph of Miller’s famous paper in which he is intent upon stimulating our interest in the magical numbers. He invites us to consider the possibility that something “deep and profound” might explain them and challenges us to prove that the similarity of the magical numbers is not simply coincidental: “I anticipate that we will find a very orderly set of relations describing what now seems an uncharted wilderness of individual differences” (*Summary*, p. 96). It is easy to share Miller’s enthusiasm for the magical numbers and the possibility that a unitary underlying capacity limit exists that constrains performance in diverse tasks. This sort of simplifying notion, a different way to look at well-known phenomena, is the stuff of important scientific discoveries (Kuhn, 1970, chaps. VI-VII, pp. 52-76; Toulmin, 1953/1960, p. 20). Quantitative constants and limits such as the gravitational constant and the speed of light have had profound influence in other sciences. It underscores Miller’s stature as a scientist that he brought the magical numbers to the attention of the field.

How far have we come in the 40-some years since the magical number paper? Shiffrin & Nosofsky (1994, p. 360) sum it up fairly accurately: “The existence of the limitations and the importance of such data for theory remain unchallenged …. If there is a deep connection, it continues to escape the field.” To that can be added, a “deep connection” has been asserted (span theory; Bachelder & Denny, 1977a, 1977b; see also Bachelder, 1997, and Denny, 1980), but that work is not well known and its basic propositions remain undebated in the literature. One of the barriers to a wider evaluation of span theory is the pervasive, prominent, and erroneous presumption that the magical numbers are unrelated and that Miller's paper bolsters that traditional point of view.

This working presumption pervades a chapter by Spitz (1973) on channel capacity in individuals with mild mental retardation. Others state the presumption more directly:

> Even though the span of absolute judgment is about seven categories, the span of attention is about six or seven objects, and the span of immediate memory is about seven items, these are not all aspects of a single underlying process. (James G. Miller, 1978, p. 137)

> A careful survey of the literature led Miller to conclude that the span of absolute judgments and the immediate memory span were unrelated. (Siegel & Siegel, 1972, p. 313)

> The possibility that there may exist some deeper connection may have contributed to the fame of this article, despite Miller's (1956) *convincing and amusing arguments* [italics added] to the contrary. (Shiffrin & Nosofsky, 1994, p. 360)

> He went on to demonstrate the *crucial difference* [italics added] between the limitations on span and on absolute judgment, with judgment being limited by the amount of *information*, measurable in bits, whereas immediate memory span is determined by the *number* of items, or to be more accurate, the number of *chunks*. (Baddeley, 1994, p. 354)

Baddeley's assertion of a "crucial difference" between absolute judgment and memory span is logically equivalent to the assertion of coincidence.

In his discussion of Miller's paper Baddeley takes pains to minimize the importance of the concept of a magical number:

> The article, if not the number seven, retains its magic. (1994, p. 356)

> The information-processing metaphor and the general utility of the concept of limited channel capacity have been enormously influential … and continue to be valuable. (1994, p. 353)
>
> As an expository *device* [italics added], I think this was brilliant, allowing Miller to link together a range of phenomena and generate what is arguably the best title in psychology, combining as it does the underlying concept of the general limit to cognitive processing capacity, with a tongue-in-cheek hint of mysticism and numerology. (1994, p. 353)

Baddeley characterizes Miller's idea of the magical number as a rhetorical device, as if it were only a bait-and-switch tactic to stimulate interest in diverse, but unrelated limitations in cognitive processing capacity. There is another way to read Miller's paper: Knowing the importance of such quantitative limits in the other sciences, Miller took the magical numbers seriously and set out to evaluate a unitary explanation of them, namely, the *channel capacity hypothesis*. He looked carefully for a deep connection, but his analyses failed. The hint of mysticism and numerology is a veil lending allure to that failure.

There are, of course, numerous examples of the opinion that the span limits are not unitary. Wickens (1984) argues for multiple capacity limits. Meyer & Kieras (1997) argue against any capacity limit at all. See Cowan (2001) for a review of these arguments.

Various links between the several spans have been proposed. Cowan (1999, 2001) has developed a variant of the working memory model which he calls "the embedded processes model of working memory." He hypothesizes that many observed performance limitations are explained by a capacity-limited attentional subsystem constraining the number of items which can be kept in mind at the same time. This concept is a general one which might well apply to all three spans (Bachelder, 2001). Cowan focusses on serial recall tasks (short-term-memory and memory span) but explicitly extends the attention concept to the span of apprehension tasks.

Siegel & Siegel (1972) reviewed the literature and compared the empirical characteristics of absolute judgment and paired-associates learning. They noted: (a) The procedures of the two tasks are almost identical. (b) Both tasks show the inverted-U-shaped error distribution (called the serial position effect in verbal learning and the end effect or bow effect in absolute judgment). and (c) The main difference between the two tasks is that the stimuli in paired-associates learning are usually, but not invariably, highly meaningful, while the stimuli in absolute judgment are of extremely low meaningfulness. They concluded:

> "The absolute judgment channel capacity and the immediate memory span for verbal material both appear to be a limitation on memory during the initial stages of practice" (p. 313).
>
> What then of the magical number seven (plus or minus two)? At least in the case of the immediate memory span and the span of absolute judgments, there may be something more involved here than just a "pernicious Pythagorean coincidence." (p. 313)

Hu (1997, p. 333), in an experimental study of absolute judgment, drew a conclusion similar to that of Siegel & Siegel (1972). Hu hypothesized that subjects keep previous trials in mind as a basis for judging the current stimulus so that the span limit in absolute judgment results from the limited working memory. "The present findings may serve as evidence of a deeper relationship between the channel capacity effect and the limit of short-term memory" (p. 333).

Killeen & Taylor (2000) make a closely related point linking memory span and absolute judgment (often called absolute identification):

> Recall of serial order [memory span] is not unlike absolute identification, as in both tasks observers must generate a one-to-one mapping of stimuli and responses.
>
> If this purported mechanism--mapping serial order onto a stochastic counter--is correct, it explains why the information transmission capability of humans is about the same for serial position recall and absolute identification.

Despite these challenges, the prevailing interpretations of the magical numbers are consistent with the traditional point of view that the magical numbers are fundamentally different phenomena, reflecting limitations in the mental faculties of attention, memory, and psychophysical judgment. Since Miller's paper is often cited to bolster this traditional interpretation, his paper and misinterpretations of it tend to undermine challenges to the traditional point of view. The purpose of the present paper is to clarify what Miller actually did in his paper and show that it is not particularly strong support for the traditional point of view let alone strong evidence against a unitary interpretation of the magical numbers. It will be shown that Miller tried to explain each of the three magical numbers with his channel capacity hypothesis. It worked reasonably well for the span of absolute judgment, but not well at all for span of apprehension, span of immediate memory, or for multidimensional absolute judgment. In other words, Miller explored one unitary explanatory hypothesis which failed. He explicitly withheld judgment about the larger issue of whether all three magical numbers reflect a unitary limitation (Miller, 1956, *Summary*, p. 96).

The Magical Numbers: Tasks and Data

This section summarizes the basic procedures and data of the three magical number tasks. Miller's presentation is quoted to make clear that the present discussion is based precisely on the same material Miller used and the way he characterized it. Figure 1 presents additional magical number data in the traditional form, that is, probability of a correct response as a function of number of stimuli. Miller introduced the magical number phenomena in this form, but then evaluated them in terms of "information" concepts and statistics. Ironically, information concepts and statistics actually obscure the fundamental similarity of the three tasks.

Figure 1 shows that number of stimuli is a highly potent independent variable in all three tasks. From about 4 through 11 stimuli the probability of a correct response changes dramatically from a high near 1.0 to a low near 0. Miller acknowledged the importance of this variable through the use of the term *stimulus information* defined as $\log_2$(number of stimuli). Number of

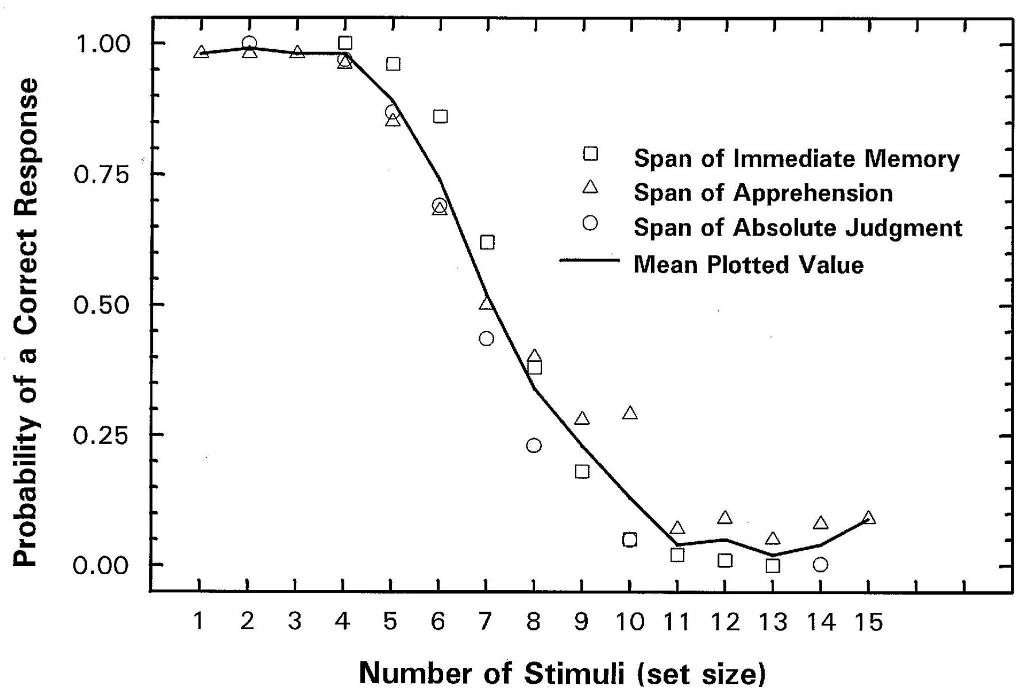

Figure 1. The probability of a correct response as a function of number of stimuli (set size) in the immediate memory, span of apprehension, and absolute judgment tasks. The data are from Guilford & Dallenbach (1925, p. 627, Table II), Mandler & Shebo (1982, p. 8, Figure 3), and Pollack (1952, p. 748, Figure 7), respectively.

stimuli is sometimes referred to as *size of the stimulus set* (e.g., LaCouture, 1997, p. 121; Lacouture & Marley, 1995, p. 384). The corresponding span theory term is *span load*, the number of items in a *relevant stimulus set*. The term span load has replaced *task complexity*, the previous span theory term introduced by Bachelder and Denny (1977a, 1977b). The term *set size* is used below for its convenience and comparative theoretical neutrality.

Another way to characterize the set size variable is that it produces a pronounced *threshold* effect. The point where performance shifts from usually right to usually wrong is often referred to as a discontinuity or threshold. The spans or magical numbers are quantifications of the thresholds in all three tasks.

The similarity of the curves in Figure 1 may surprise even specialists in span phenomena because the three types of span task are not usually measured in directly comparable ways so their fundamental similarity is obscured. Were the curves in Figure 1 not labeled, the reader could easily presume they all plot minor variations in the same task in the same laboratory in the same study. In fact, these data span 57 years (1925 to 1982), are from different laboratories with different underlying theoretical approaches, and derive from rather different-looking tasks which traditionally have been considered fundamentally distinct. The following paragraphs present the procedures for the three span tasks.

Span of immediate memory. In the most common version of this task the participants hear several spoken words or digits then attempt to respond with a series of responses, each corresponding to a single stimulus. “Everybody knows that there is a finite span of immediate memory and that for a lot of different kinds of test materials this span is about seven items in length” (Miller, 1956, p. 91). When it is said that memory span is about seven items in length, what is meant at the data level is that college students, repeating digit or word sequences, usually make no errors on trials with 1 through 5 or 6 items, err in about 50% of trials with 6 or 7 items, and err on nearly all trials with 8 or more items. Miller did not cite a specific study to support his characterization of immediate memory span, but Figure 1 presents representative data by Guilford & Dallenbach (1925, p. 627, Table II). They presented stimulus sequences ranging from 4 to 13 digits, one of each size, to 100 different subjects. The subjects, though not specified, almost certainly were college students.

Span of apprehension. In this task the stimuli are random patterns of varying numbers of items such as dots, squares, circles, or triangles flashed rapidly and simultaneously on a screen. The observer responds by reporting the number of items. Presentation is slow enough to be clear what was shown, but fast enough that the observer cannot count in the usual sense of counting. Miller based his discussion on reference experiments by Kaufman, Lord, Reese, and Volkmann

(1949) who presented 1 to 210 dots at an exposure of 1/5 sec.

> The first point to note is that on patterns containing up to five or six dots the subjects simply did not make errors. The performance on these small numbers of dots was so different from the performance with more dots that it was given a special name. Below seven the subjects were said to *subitize*; above seven they were said to *estimate*. (Miller, 1956, p. 90)

Figure 1 presents span of apprehension data taken directly from Mandler & Shebo (1982, p. 8, Figure 3). The only changes made were to convert from probability of an error to probability of a correct response by subtracting Mandler & Shebo's values from 1.0. These data were selected for Figure 1, because the study Miller cited (Kaufman, et al., 1949) used a very different dependant variable, namely, median number of dots reported. On small set sizes (1 to about 5 or 6 items) the values of both variables are virtually identical (compare Figure 1, here, with Figure 6 and Table I of Kaufman, et al., 1949). For larger set sizes the two variables diverge sharply. Probability of a correct response drops rapidly to near 0 as shown in Figure 1; but median number of dots reported increases almost linearly with set size.

Span of absolute judgment for unidimensional stimuli. Both Miller's presentation and Figure 1 draw upon the same data (Pollack, 1952, p. 748, Figure 7). In this task the items of a stimulus set vary in magnitude along one perceptual dimension such as size, pitch, loudness, or brightness. The procedure tests members of a stimulus set one by one and the subject responds with a number indicating the rank of the item within the full stimulus set. The task is often called absolute identification even though the judgments are clearly relative, not absolute (Garner, 1962, p. 77).

> [Pollack] asked listeners to identify tones by assigning numerals to them. The tones were different with respect to frequency, and covered the range from 100 to 8000 cps in equal logarithmic steps. A tone was sounded and the listener responded by giving a numeral.

> When only two or three tones were used the listeners never confused them. With four different tones confusions were quite rare, but with five or more tones confusions were frequent. With fourteen different tones the listeners made many mistakes. (Miller, 1956, p. 83)

Pollack's data had to be corrected to be metrically consistent with the other two curves of Figure 1. This is because both tasks, span of immediate memory and span of apprehension, present the full stimulus set on each trial, while span of absolute judgment tasks present a single stimulus on each trial. For example, in a traditional span of immediate memory task when *S* hears 7 digits (a full stimulus set) and responds perfectly the response is scored "correct" for the full stimulus set considered collectively. Any error, even a simple omission of a single digit, is recorded as "wrong" for the full stimulus set. Similarly, in a span of apprehension task *S* sees the full stimulus set, responds with a number, and is scored either "right" or "wrong" with respect to the full stimulus set considered collectively. Span of absolute judgment, however, is usually scored either in terms of the information metric, I_t (which summarizes responding to the full stimulus set considered collectively), or as probability of a correct response to a single stimulus within a full stimulus set.

Pollack presented the same data set both ways, as I_t and as probability of a correct response to a single stimulus (his Figures 3 and 7, respectively). The estimated probability of correctly judging all stimuli of a full stimulus set is simply the product of the individual probabilities, namely, $p' = p^{\text{set size}}$. For example, Pollack's Figure 7 shows a mean probability of about .94 for a correct response to a single stimulus for set size 6. The estimated probability of performing perfectly on the full stimulus set considered collectively is $.94^6$ which equals .69. The corrected values are plotted in Figure 1.

The span data in subjects with retardation are highly consistent with the notion the span limits are unitary and covary closely with each other and with intelligence (for a review of these data see Bachelder & Denny, 1977a, 1977b). Spitz (1973) reviewed spans in subjects with mental retardation and found that all three spans are lower than those of non-retarded subjects, 5 ± 2 rather than 7 ± 2. Bachelder (1976; summarized in Bachelder & Denny, 1977a, p. 139-142) measured both span of immediate memory and span of absolute judgment in 60 adults: 20 college students (above average intelligence), 20 institutional cottage staff (average intelligence), and 20 institutional residents with retardation (mean IQ=47.8). The correlation between span of immediate memory (measured as the traditional threshold) and span of absolute judgment (measured as I_t) was .78.

Bachelder plotted the same performances by the same subjects in a different way which emphasizes not only the close covariation of the two spans, but the strong similarity in their numerical values. He defined span of absolute judgment, not as I_t, but as the largest set size for which performance was 90% correct or higher. This criterion for single stimuli corresponds closely with the criterion of 50% correct for all stimuli of a set considered collectively. For example, for set size 5, $.90^5 = .59$.

Bachelder regrouped his 60 subjects (ignoring their status as retarded, average, or above average in intelligence) into relatively homogenous groups on the basis of immediate memory span for common single-syllable words. He then plotted mean span of absolute judgment as a function of mean span of immediate memory. Those data are presented here as Figure 2. The figure shows that mean span of absolute judgment not only covaries closely with mean span of immediate memory, but is essentially identical in value.

One of the reasons it has been difficult to accept the idea of a unitary magical number is the fairly wide range of published estimates of the value. Figure 1 shows clearly that the value one takes for the magical number must vary greatly with the criterion one uses to define it. To read the magical numbers from Figure 1, simply inspect the figure to determine the set size which corresponds to 50% correct, the traditional definition of span threshold. In Figure 1 the mean performance for set size 7 falls almost precisely at the 50% point. This is Miller’s magical number 7. In contrast, Broadbent (1975) and Cowan (2001) estimated the magical number to be

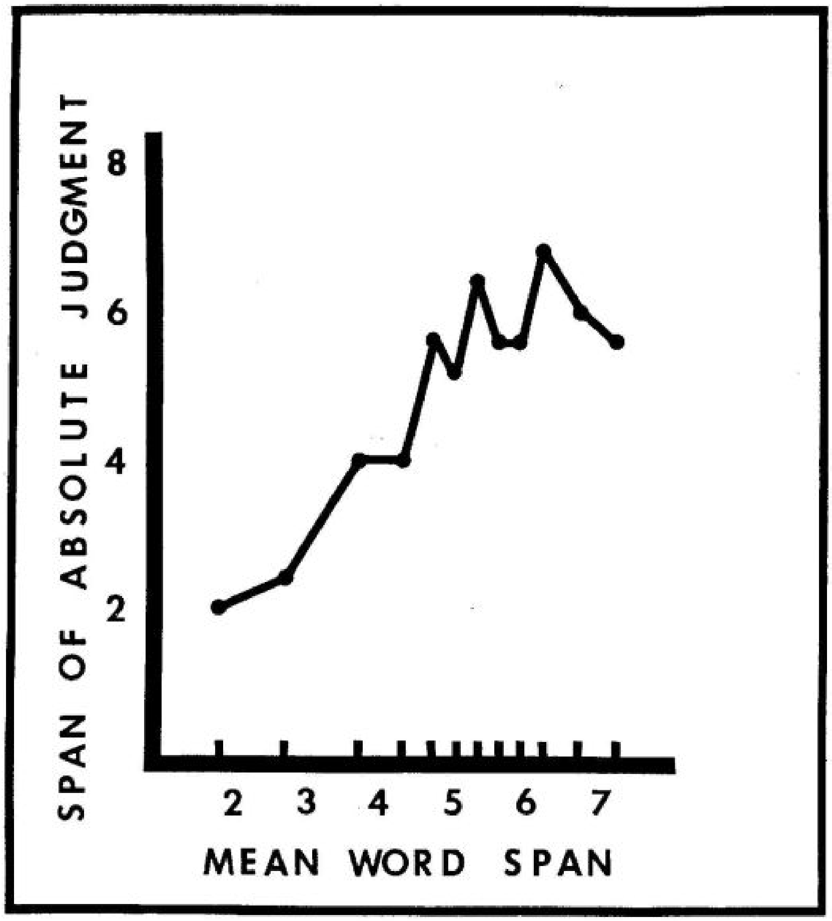

Figure 2. The span of absolute judgment as a function of span of immediate memory (for familiar, single-syllable words) in bright normal, normal, and retarded (IQ=47.8) subjects. Each point represents five subjects of relatively homogeneous psychometric span ability (disregarding IQ status). (From "A theory of intelligence: I. Span and the complexity of stimulus control" by Bruce L. Bachelder and M. Ray Denny, *Intelligence*, 1(2), p. 141. Copyright 1977 by Ablex Publishing Corporation. Reprinted by permission. [Permission not yet requested.])

about 3 and 4, respectively; but they explicitly used a 100% criterion. Figure 1 shows that a performance criterion of 100% correct puts the magical number at 4 ± 1. From the perspective of Figure 1 there is no fundamental difference between Miller's 7, Broadbent's 3, and Cowan's 4; all are simply different points on the same curve relating probability of a correct response to set size.

Miller's Theoretical Analyses

This section summarizes Miller's theoretical analyses of the three span tasks and multidimensional absolute judgment. Psychometric concepts and terminology will help explicate what Miller actually did. The psychometric tradition is based on *classical test theory* which asserts that a psychological test measures a *latent variable*, that is, a psychological construct which is not directly observable. Examples of such latent variables include: intelligence, working memory, attention, and span ability. The term *face validity* "refers, not to what the test actually measures, but to what it appears superficially to measure" (Anastasi, 1982, p. 136). Statements of face validity are hypotheses, judgments, and conventions; they must be tested empirically (Cronbach, 1957, p. 676). Testing face validity hypotheses is called *construct validity* research. A test has construct validity to the extent it behaves in accordance with our theory about the construct. The concept of construct validity was introduced by Cronbach & Meehl (1955). See Anastasi (1982), Messick (1989), or Standards (1985) for more recent discussions of validity.

In the paragraphs below it will be shown that Miller asserted a new face validity hypothesis, the channel capacity hypothesis, as an alternative to the traditional face validities of the three span tasks. For each task in turn he evaluated the construct validity of the new channel capacity validity hypothesis. His analyses supported the hypothesis in the case of span of absolute judgment, but not in the case of either span of apprehension, span of immediate memory, or multidimensional absolute judgment.

Span of absolute judgment. Recall that in absolute judgment tasks the subject encounters stimuli one at a time and responds with a number naming the rank of the stimulus in the stimulus set. When set size is small, performance is near perfect; but as set size increases, a point of discontinuity is reached; see Figure 1 (Pollack's data). The following lines show how Miller reinterpreted absolute judgment tasks as measures of channel capacity.

> In the traditional language of psychology these would be called experiments in absolute judgment. Historical accident, however, has decreed that they should have another name. We now call them experiments on the capacity of people to transmit information. (p. 81)
>
> In the experiments on absolute judgment, the observer is considered to be a communication channel. (p. 82)
>
> If the human observer is a reasonable kind of communication system, then when we increase the amount of input information the transmitted information will increase at first and will eventually level off at some asymptotic value. This asymptotic value we take to be the *channel capacity* of the observer. (p. 82)

The statistic which defines the concept of channel capacity is I_t, a complex statistic measuring the extent of covariation of responses and stimuli ("measure of the input-output correlation," Miller, 1956, p. 82). See Garner & Hake (1951) for the details of calculation. Miller (p. 86) took "the best estimates" of I_t for a wide range of types of stimuli in absolute judgment experiments and found a mean of about 6.5 "categories" of judgment for the channel capacities of college students. This value is rather similar to the magical numbers in span of apprehension and span of immediate memory. Miller must have been greatly encouraged that his channel capacity hypothesis would be fruitful because of the simple mathematical relation between I_t and traditional span thresholds and because the plots of I_t versus stimulus information ($\log_2$[set size]) behave remarkably like analogous plots from the communications engineering laboratory. Thus, his analyses bolster, for absolute judgment tasks, the construct validity of the concept of channel capacity defined by I_t.

Span of apprehension. Recall that in span of apprehension observers see brief presentations of groups of items varying in number and respond with the number of items shown. See Figure 1 (data by Mandler & Shebo). Miller explored the hypothesis that the magical number in span of apprehension is due to a limited channel capacity:

> This is, as you will recognize, what we once optimistically called "the span of attention." (Miller, 1956, p. 90)
>
> This discontinuity at seven is, of course, suggestive. Is this the same basic process that limits our unidimensional judgments to about seven categories? The generalization is tempting, but not sound in my opinion … . on the basis of the published data I would guess that the subjects transmitted something more than four bits of information about the number of dots … . This is considerably more information than we would expect to get from a unidimensional display.

Note how Miller alludes to "attention," the traditional face validity of this task. He is alerting the reader that he will assert a new validity hypothesis, the channel capacity hypothesis. In the lines above Miller "tested" then rejected his channel capacity hypothesis as an explanation of the magical number in span of apprehension tasks.

Span of immediate memory. In an argument which parallels his arguments with respect to the spans of absolute judgment and apprehension, Miller evaluated the hypothesis that the underlying limitation in span of immediate memory is a limitation in channel capacity:

> Everybody knows that there is a finite span of immediate memory and that for a lot of different kinds of test materials this span is about seven items in length. I have just shown you that there is a span of absolute judgment that can distinguish about seven categories and that there is a span of attention that will encompass about six objects at a glance. What is more natural than to think that all three of these spans are different aspects of a single underlying process? And that is a fundamental mistake, as I shall be at some pains to demonstrate.
>
> My mistake went something like this. We have seen that the invariant feature in

> the span of absolute judgment is the amount of information that the observer can transmit … . If immediate memory is like absolute judgment, then it should follow that the invariant feature in the span of immediate memory is also the amount of information that an observer can retain. If the amount of information in the span of immediate memory is a constant, then the span should be short when the individual items contain a lot of information and the span should be long when the items contain little information. (p. 91)

Miller then cited a study by Hayes (1952) indicating that spans for immediate memory are all essentially the same whether the stimuli are drawn from pools of two stimuli (binary digits, low information), 10 stimuli (the digits 0-9, intermediate information), or 1000 stimuli (1000 common words, high information). At this point he asserted the magical numbers are coincidental: “In spite of the coincidence that the magical number seven appears in both places, the span of absolute judgment and the span of immediate memory are quite different kinds of limitations that are imposed on our ability to process information” (p. 92).

Thus, Miller rejects his channel capacity hypothesis as an explanation of the magical number in memory span. He then asserted (p. 92) a new validity hypothesis, the *chunking* hypothesis, to explain the magical number limit in span of immediate memory.

Multidimensional absolute judgment. In this type of task the stimuli vary along more than one dimension and Miller (pp. 87-89) reported that measures of information transmission greatly exceed those obtained in unidimensional absolute judgment. So in a way, the data of multidimensional absolute judgment tasks provide a *coup de grace* for Miller’s channel capacity hypothesis as a unitary explanation of the several spans.

In summary, taking Miller’s analyses at face value, it can be seen that the channel capacity hypothesis is valid only for the absolute judgment tasks. Miller’s analyses reject the hypothesis as a unitary explanation of the several limits; but another unitary hypothesis might well explain all three limits and thereby strongly challenge the prevailing presumption of coincidence. A challenge to conventional wisdom would be bolstered by tenable alternative interpretations of the magical numbers. Papers by Cowan (1999, 2001; also see Bachelder, 2001) suggest a unitary interpretation in terms of a capacity-limited central attention concept. Bachelder and Denny (1977a, 1977b) assert a unitary explanation in terms of the span ability concept. Span theory is comparatively unknown and fundamentally different from the cognitive approach to this issue so a sketch of the theory is included below.

Span Theory

Span theory is an unusual theory in that it combines key concepts from the psychometric tradition and the neobehavioral S-R tradition in an attempt to understand “higher cognitive function,” long the province of the cognitive tradition. Span theory asserts the working hypothesis that the three magical numbers all reflect a unitary limitation in *span ability*, an individual differences construct (latent variable) assessed via an immediate memory span test. The term span ability was chosen originally for convenience and theoretical neutrality to refer, simply, to that which is measured by span tests, avoiding a premature commitment to a specific validity hypothesis. The research strategy relies on construct validity research to determine the theoretical nature of span ability.

The first validity hypothesis in the span theory research program was the “elicitation span” hypothesis (Bachelder, 1970/1971, p. 2). According to this hypothesis, the span limitation in span of immediate memory tasks is a limitation on the number of stimuli which can elicit a sequence of responses called the “response string.” This face validity hypothesis is more akin to the S-R behavioral concept of stimulus control ($S_1S_2...S_n$-$R_1R_2...R_n$) than to the cognitive concept of memory. Bachelder (1977) summarized his research which explored a closely related validity hypothesis, namely, that span ability is the ability to imitate verbal sequences and language. In a variety of experiments with non-retarded and retarded subjects he found that individual differences in the ability to repeat sentences covaries closely with psychometric span ability (immediate memory span). In psychometric terms, the test of span ability has strong construct validity as a measure of the ability to imitate sentences and other language sequences.

The concepts of elicitation span and language imitation ability were and continue to be useful ways to view the span limitation in serial recall tasks and language imitation tasks; but they are not general enough. They do not link the span of immediate memory with the spans of apprehension and absolute judgment because these latter two tasks do not have extended stimulus and response sequences as do serial recall tasks and language imitation tasks.

The concept linking all three spans, the *joint relevance hypothesis,* was introduced by Bachelder and Denny (1977a, 1977b). They hypothesized that span ability is the ability to function under complex stimulus control, defined as situations in which two or more stimuli are jointly (conjunctively) relevant for target responding. Complexity is equal to the number of jointly relevant stimuli. In lay terms, joint relevance refers to situations in which correct performance requires that two or more stimuli be kept in mind at the same time. In recent years, the term span load has replaced the term complex stimulus control.

In the three magical number tasks set size corresponds to span load. The empirical curves plotting probability of a correct response as a function of span load are inverse ogives as shown in Figure 1. Performance under low span loads (1 to about 4 or 5 in college students) is near perfect; performance at higher span loads declines abruptly and markedly. The threshold, the point where performance is 50% correct, is the measure and definition of psychometric span ability. Thus, according to span theory, performance in diverse tasks is an inverse ogival function of span load. This mathematical formulation, an empirical generalization, explains the three span limits shown in Figure 1 as well as numerous other data (Bachelder & Denny, 1977a, 1977b). Note that Miller's descriptions of the data of the three span tasks correspond closely with this mathematical form.

Span theory also offers an alternative interpretation of multidimensional absolute judgment data which were an important part of Miller's presentation. Recall that the multidimensional absolute judgment tasks provided something of a coup de grace for the unitary channel capacity hypothesis because information transmitted in these tasks greatly exceeds the magical number 7. Miller illustrated his arguments with a study by Pollack & Ficks (1954). Span theory analyses are able to explain these same multidimensional data while retaining the notion of a unitary limitation in span ability having the same value as in the three basic span tasks of Figure 1.

Bachelder (1978, Part 2) translated Pollack & Ficks' data from information terms back to probabilities of a correct response and modeled their task as a combination of serial recall and unidimensional absolute judgment: subjects judged the values in each of multiple dimensions then retained and reported the values as in serial recall. Mathematical models, presuming unitary magical number limitations typical for college students, generate the published data remarkably precisely. The mean error is 2.7%. Seven of the nine predicted values fall within 12% of the published values. The two poorest predictions are 26.3% low and 17.0% high. In other words, span theory models, presuming span ability values in the usual 7 ± 2 range, explain the high values of information transmission in multidimensional absolute judgment. Thus, Pollack & Ficks' data are fully consistent with the hypothesis of a unitary limitation acting, not only in the three magical number tasks, but also in a multidimensional absolute judgment task.

Summary and Conclusions

It is widely believed that Miller proved the magical numbers are coincidental, but Miller himself was more temperate in his conclusions. In his *Summary* he "proposed to withhold judgment" and "suspected coincidence." Miller's temperate tone is fitting, since a conclusion of coincidence does not follow from his arguments.

Miller challenged the traditional interpretation of the three magical numbers as fundamentally different (measures of memory, judgment, and attention). He "tested" his channel capacity hypothesis as a unitary explanation of the magical numbers. He found the hypothesis is nicely supported in span of absolute judgment of unidimensional stimuli but not at all in span of apprehension, span of immediate memory, or absolute judgment tasks with multidimensional stimuli. When his channel capacity hypothesis failed to explain the magical number in span of immediate memory, he challenged tradition again by asserting the chunking hypothesis to explain the magical number limitation in serial recall tasks.

The failure of Miller's analyses did not prove the magical numbers are unrelated, but only that the channel capacity hypothesis, as Miller tested it, fails to explain all of them. Numerous data are consistent with a unitary explanation (e.g., Bachelder, 1976, summarized in Bachelder & Denny, 1977a, p. 139-142; Guilford & Dallenbach, 1925; Mandler & Shebo, 1982; Pollack, 1952; and Spitz, 1973). For reviews of additional data of this sort see Bachelder and Denny (1977a, 1977b) and Cowan (2001).

Miller rejected the channel capacity hypothesis, but a different unitary hypothesis might explain the limitations. Current unitary explanatory hypotheses include span theory (Bachelder, 1997; Bachelder & Denny, 1977a, 1977b; Denny, 1980), the attention hypothesis of the embedded processes working memory model (Cowan, 1999; 2001), and the signal detection model of discrimination (Killeen & Taylor, 2000).

One of the reasons it has been so easy to accept Miller's paper as proof of coincidence is the long tradition of viewing the span tasks as fundamentally different--as measures of judgment, attention, and memory. As Kuhn (1970, p. 77) pointed out, adherence to such traditional working presumptions is difficult to change in a discipline at large, even if the presumptions clearly conflict with valid data.

Interpretation of the three magical numbers as limitations on judgment, attention, memory, working memory, chunking, or the ability to cope with span load are face validities, hypotheses based on appearance, professional judgment, and convention. While there is a good deal of research based on one or another face validity hypothesis, there is comparatively little systematic research which pits one validity hypothesis against another in an attempt to determine the nature of the limits in the several tasks.

The necessity for construct validity research is more in the psychometric than in the experimental tradition. Miller's (1956) paper came out just a year after Cronbach & Meehl

(1955) introduced the concept of construct validity and a year before Cronbach's (1957) famous call for a rapproachment between the "correlational" and "experimental" psychologies. Psychometrics was in the early years of development of validity concepts about the time Miller was working on his "magical number" paper. This is probably why Miller was able to assert new validity hypotheses with little challenge to justify them. There simply was no body of literature to "keep him honest," so to speak, and no tradition in experimental psychology requiring that assertions of face validity be evaluated formally.

Another reason it has been easy to accept Miller's paper as proof of coincidence is the more recent tradition of research on the three types of tasks. Except for span theory research, which approaches all three tasks with the same basic behavioral-style theory, quite different sorts of theory are used to investigate the three tasks. In general, the immediate memory (serial recall) tasks are investigated with variations of Baddeley's (1986) working memory model. Absolute judgment (identification) is studied with quite different models, such as the theory of auditory intensity resolution (Durlach & Braida, 1969) or the attention band theory (Luce, Green, & Weber, 1976). Span of apprehension (attention) is studied with yet different models, e.g., the template matching model of Klahr (1973a, 1973b) or the FINST (finger of instantiation) theory of Trick & Pylyshyn (1994). These several models are very different so it is "natural" to presume the limits are fundamentally different. The span theory model applies to all three tasks, so it is "natural" from a span theory perspective to hypothesize the tasks are fundamentally the same.

The magical numbers are important and still need to be explained, a point also asserted by Cowan (1999, 2001) and Shiffrin & Nosofsky (1994, p. 360). The application of the concepts and methods of psychometrics, that is, construct validity research involving the assertion and experimental test of validity hypotheses is a potent tool in this endeavor. This writer is inclined to believe the span ability construct explains the span limitations well, but that is a matter for continued debate and validity research.

References

Anastasi, A. (1982). *Psychological testing, Fifth Edition.* New York: Macmillan.

Bachelder, B. L. (1970/1971). The memory span paradigm: Its use for the analysis of mental retardation. (Doctoral Dissertation, Michigan State University, 1970). *Dissertation Abstracts International, 32*(1), 576B-577B.

Bachelder, B. L. (1976). Memory span and absolute judgment in normal and retarded adults. Unpublished manuscript.

Bachelder, B. L. (1977). Memory span tests and the ability to imitate word sequences and sentences. *Ohio Journal of Speech and Hearing, 13*(1), 5-15. Retrieved from http://www.brucebachelderphd.com/SpanAndLanguage.pdf

Bachelder, B. L. (1978, March). Span Theory: Laboratory and classroom applications. Paper presented at the Eleventh Annual Gatlinburg Conference on Research in Mental Retardation, Gatlinburg, TN. Retrieved from http://www.brucebachelderphd.com/Bachelder_1978_Gatlinburg.pdf

Bachelder, B. L. (1997, May 17). An overview of span theory: A poster session presented at The Golden Anniversary Celebration, Department of Psychology, Michigan State University. E. Lansing, MI.

Bachelder, B. L. (2001). The magical number 4 = 7: Span theory on capacity limitations. *Behavioral and Brain Sciences, 24*(1), 116-117.

Bachelder, B. L., & Denny, M. R. (1977a). A theory of intelligence: I. Span and the complexity of stimulus control. *Intelligence, 1*(2), 127-150.

Bachelder, B. L., & Denny, M. R. (1977b). A theory of intelligence: II. The role of span in a variety of intellectual tasks. *Intelligence, 1*(3), 237-256.

Baddeley, A. D. (1986). *Working memory.* Oxford: Clarendon Press.

Baddeley, A. (1994). The magical number seven: Still magic after all these years? *Psychological Review, 101*(2), 353-356.

Broadbent, D. E. (1975). The magic number seven after fifteen years. In A. Kennedy & A. Wilkes (Eds.), *Studies in long term memory* (pp. 3-18). London: Wiley.

Cowan, N. (1999). An embedded-processes model of working memory. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 62-101). Cambridge, UK: Cambridge University Press.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24,* 87-185.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *The American Psychologist, 12,* 671-684.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281-302.

Denny, M. R. (1980). Complex behavior: Traditional comparative psychology. In M. R. Denny (Ed.), *Comparative psychology: An evolutionary analysis of animal behavior* (pp. 249-266). New York: John Wiley.

Durlach, N. I., & Braida, L. (1969). Intensity perception. I. Preliminary theory of intensity resolution. *Journal of the Acoustical Society of America, 46,* 372-383.

Garner, W. R. (1962). *Uncertainty and structure as psychological concepts.* New York: Wiley.

Garner, W. R., & Hake, H. W. (1951). The amount of information in absolute judgments. *Psychological Review, 58,* 446-459.

Guilford, J. P., & Dallenbach, K. M. (1925). The determination of memory span by the method of constant stimuli. *American Journal of Psychology, 36,* 621-628.

Hayes, J. R. M. (1952, January-June). Memory span for several vocabularies as a function of vocabulary size. In *Quarterly Progress Report.* Cambridge, MA: Acoustics

Laboratory, Massachusetts Institute of Technology.

Hu, G. (1997). Why is it difficult to learn absolute judgment tasks? *Perceptual and Motor Skills, 84*(1), 323-335.

Kaufman, E. L., Lord, M. W., Reese, T. W., & Volkmann, J. (1949). The discrimination of visual number. *The American Journal of Psychology, 62,* 498-525.

Killeen, P. R., & Taylor, T. (2000). How the propagation of error through stochastic counters affects time discrimination and other psychophysical judgments. *Psychological Review, 107*(3), 430-459.

Klahr, D. (1973a). An information processing approach to the study of cognitive development. In A. D. Pick (Ed.), *Minnesota Symposia on Child Psychology.* (Vol. 7). Minneapolis: University of Minnesota Press.

Klahr, D. (1973b). A production system for counting, subitizing and adding. In W. G. Chase (Ed.), *Visual information processing.* New York: Academic Press.

Kuhn, T. S. (1970). *The structure of scientific revolutions* (2nd ed.). Chicago: University of Chicago Press.

Lacouture, Y. (1997). Bow, range, and sequential effects in absolute identification: A response-time analysis. *Psychological Research, 60,* 121-133.

Lacouture, Y., & Marley, A. A. J. (1995). A mapping model of bow effects in absolute identification. *Journal of Mathematical Psychology, 39*(4), 383-395.

Luce, R. D., Green, D., & Weber, D. (1976). Attention bands in absolute identification. *Perception & Psychophysics, 20,* 49-54.

Mandler, G., & Shebo, B. J. (1982). Subitizing: An analysis of its component processes. *Journal of Experimental Psychology: General, 111*(1), 1-22.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (pp. 13-58). New York: American Council on Education/Macmillan.

Meyer, D., & Kieras, D. (1997). A computational theory of executive processes and multiple-task performance: Part 1. Basic mechanisms. *Psychological Review, 104,* 3-65.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review, 63,* 81-97.

Miller, J. G. (1978). *Living systems.* New York: McGraw-Hill.

Pollack, I. (1952). The information of elementary auditory displays. *Journal of the Acoustical Society of America, 24,* 745-749.

Pollack, I., & Ficks, L. (1954). Information of elementary multi-dimensional auditory displays. *Journal of the Acoustical Society of America, 26,* 155-158.

Shiffrin, R. M., & Nosofsky, R. M. (1994). Seven plus or minus two: A commmentary on capacity limitations. *Psychological Review, 101*(2), 357-361.

Siegel, J. A., & Siegel, W. (1972). Absolute judgment and paired-associate learning: Kissing cousins or identical twins? *Psychological Review, 79,* 300-316.

Spitz, H. H. (1973). The channel capacity of educable mental retardates. In D. K. Routh (Ed.), *The experimental psychology of mental retardation* (pp. 133-156). Chicago: Aldine.

Standards. (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Toulmin, S. (1960). *The philosophy of science: An introduction.* Harper Torchbooks/The Science Library. New York: Harper & Row. (Original work published 1953)

Trick, L., & Pylyshyn, Z. (1994). Why are small and large numbers enumerated differently? A limited-capacity preattentive stage in vision. *Psychological Review, 101,* 80-102.

Wickens, C. D. (1984). Processing resources in attention. In R. Parasuraman & D. Davies (Eds.), *Varieties of attention* (pp. 63-102). New York: Academic Press.

Author Note

Bruce L. Bachelder, psychologist in independent practice (retired, 2004).

I wish to thank several people for their help with this manuscript. Several critical reviews by M. Ray Denny have been especially helpful as were reviews by Nelson Cowan of earlier versions of the paper. Robert B. Freeman, Jr. and Lester M. Hyman made useful comments on an earlier version.

This article was originally prepared and submitted for publication in 2000. As of September 2015 it had not been published. Minor adjustments were made in September 2015. Send correspondence to Bruce L. Bachelder, 306 West Union Street, Morganton, NC 28655-3729 or bruce@brucebachelderphd.com